University of Reading

Essays on Sentiment: An analysis of the Commercial Real Estate Market

A thesis submitted for the degree of Doctor of Philosophy

Real Estate and Planning
Henley Business School
University of Reading

Steffen Heinig

October 2018

DECLARATION OF ORIGINAL AUTHORSHIP

I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Steffen Heinig

ACKNOWLEDGEMENTS

To my beloved wife Ginny, which has suffered an immense pain while I have pursued my dream. I am grateful for her continuous support and encouragement. I am further thankful for the experience we had over the cause of my PhD. This has made us stronger and brought us closer together as ever before. A life without you does not make much sense, and I am looking forward to the new adventures there to come. Thank you for all of this, because without you I wouldn't have been able to do any of this.

Secondly, I am grateful to my supervisor Prof. Dr. Anupam Nanda who's door was always open for urgent and not so urgent questions. Both his work and life-related advice have helped me to complete my PhD. I will always be thankful for his support, and I will wait for this chess match which hasn't been played until today.

A very special gratitude to my second supervisor Prof. Dr. Sotiris Tsolacos. Sometimes seeing the devil in person brings people together and says more than 1,000 words. Hope to catch another foamy latte with him whenever he is around.

I am also grateful to my parents Beate and Ralf Heinig, who have brought my brother and me up in the way to always think first and act second. Never accept an opinion without questioning it. I know it has been hard, but I am thankful for the support and that they made it possible for me to study abroad. Further, to my brother Jan, who is proud of my achievements while I am proud of his - without ever saying too much about this.

With a special mention to Prof. Dr. Simon Stevenson, Dr. Ranoua Bouchouicha and Jane Batchelor who have all pointed out to me that an academic career would be more suitable for my temperament and questioning skills. Without them, I would never have thought about it. I am thankful for your support throughout all the years.

Out of chronological reasons I have to mention the MSc Real Estate Finance and Investment course of 2013/14. Especially, Tom, Junda, Sophie, Candice, Lien, Jane and Winnie, you people had made Reading special to me in the first place. I will never forget our trip to Hong Kong and Taiwan.

I also like to thank those people who have successfully survived room 118. I have met the most interesting people from around the world, and I have made some incredible friends over the years. A special thanks to our little dinner group Dr. Chis Foye, Dr. Dongwoo Hyun, Dr. Shuai Shi, Dr. Mark Dobson, Dr. Xueying Chen and Joseph Ayitio (soon to be honoured with a Doctors

degree). I share with all of them some incredible memories, which will be brought back once we will meet again.

I like to specially mention Dongwoo, who has become one of the best friends I ever had. Thank you for all the moments we shared. I am stunned by your self-discipline and how you pursue your dream. Every time I have a coffee I will remember all of this - thank you.

Also, to Alex, with whom I shared an apartment for a year in the noisiest student hall on the entire campus. I have never met a person who goes so easily through life and achieves his goals with so much ease. I wish you all the best for the future.

Further to Mark, who is actually not from Reading but from a small town called Twyford. I have never met a person who loves to talk so much about his topic. Picking up where we have left is a great present and allows little maintenance.

I like also to specially mention Chris Foye, who once joked, that this acknowledgement section is going to be longer than my entire thesis. I am impressed by your wisdom, and I am thankful for many pieces of advice you gave me throughout the time we had.

Besides I like to mention Abdullah Alfalah who has shown me that family life is more important than anything else in this world. A lesson worth teaching. Till today I think about Kuwait and its warm and friendly people - I will visit it again.

There is another PhD which is worth mentioning - Filipe Morais. Even though not from Real Estate he has become a great fountain of inspiration both academically and life wise. I am thankful for so many espressi paid by the European Union.

Finally, I like to thank Helen for all the support she provided. It's a tragedy we have become friends so late, but I am thankful it happened in the end. Tea is and will always be a great measure of conversational flow. Take good care and enjoy the bright sides of life. Take also good care of Jorn, who did actually run through half of Warsaw just to get you a more suitable souvenir.

Last but not lease a great thanks to all the people I have met over the years. My time at Reading was made enjoyable in large part due to all of you.

ABSTRACT

This thesis deals with the extraction, construction and analysis of commercial real estate (CRE) sentiment within Europe and the U.K. especially. The three empirical studies in this thesis may contribute to our understanding of the discipline. As I establish in the literature review, the analysis of commercial real estate sentiment still offers a lot of potential for further research. Since real estate markets are subject to sentiment swings, scholars and market participants should consider them in their market analysis.

The first study establishes the need for sentiment consideration within the European real estate market. In order to justify the research of sentiment analysis, I have used different indirect and direct sentiment proxies and applied them in yield models for 80 different commercial property (sub-)markets within Europe. The statistical modification of different sentiment proxies is needed since not all European property markets offer direct sentiment measures. The results suggest, that the consideration of sentiment in a yield model framework adds significant information. I found, that CRE markets, which are assumed to be more liquid and developed, show a larger exposure to property specific sentiment measures. Markets, which are assumed to be less developed (i.e. Eastern European markets) on the other hand, have a larger exposure to more general macroeconomic sentiment indicators.

The second study introduces a new method, which can be used to extract sentiment from text documents. The primary motivation for the use of text documents and the application of Natural Language Processing (NLP) methods lies in the fact that these documents are published much faster than other sentiment proxies. This allows extracting a much more accurate market sentiment. The second study should be understood as an introductory chapter to the method and the field of NLP. In total four different wordlists (AFINN, BING, NRC and TM) are used to extract the sentiment form various market reports for the CRE market in U.K. The study reveals that sentiment extracted from those documents, can be used to improve autocorrelated models.

The last study uses those findings and applies different supervised learning methods. While the second study has produced sufficient results, the underlying text corpus of market reports has shown a series of insufficiencies. I have therefore, used a large dataset of more than 120,000 news articles, all concerning the British CRE market. Findings suggest, that the main issue of supervised learning algorithms is the appropriate classification of the different entities. I offer two approaches in order to construct robust sentiment indicators.

TABLE OF CONTENTS: OVERVIEW

1 INTRODUCTION 1

2 LITERATURE REVIEW 22

3 SENTIMENT PROXIES 38

4 NATURAL LANGUAGE PROCESSING 119

5 MACHINE LEARNING APPLICATION 162

6 CONCLUSION 353

7 REFERENCES 367

8 APPENDIX i

TABLE OF CONTENT

1 INTRODUCTION 1

1.1 Background and motivation 1

1.2 Aims and objectives 6

1.3 Behavioural finance origins 8

1.3.1 Behavioural finance in real estate 15

1.4 Chapter description 20

2 LITERATURE REVIEW 22

2.1 Sentiment analysis 22

2.1.1 Sentiment analysis origins 22

2.1.2 Sentiment analysis in the real estate market 30

3 SENTIMENT PROXIES 38

3.1 Introduction 38

3.2 Literature review on yield modelling 41

3.3 Theory 42

3.4 Methodology 44

3.4.1 Yield model 44

3.4.2 Sentiment measures 45

3.4.2.1 Macroeconomic sentiment indicator 46

3.4.2.2 Real estate specific sentiment indicators 48

3.4.2.3 Sentiment construction 51

3.4.2.3.1 Principal component sentiment indicators 51

3.4.2.3.2 Orthogonalization 52

3.4.2.3.3 Macroeconomic sentiment 54

3.4.2.3.4 Macroeconomic sentiment: Kaiser Criterion and PCA only 60

3.4.2.3.5 Office specific sentiment 61

3.4.2.3.6 Retail specific sentiment 65

3.4.2.3.7 Property specific sentiment 66

3.4.2.3.8 Google Trends 67

3.4.3 Empirical models 69

3.5 Data description 71

3.5.1 Google Trends data 72

3.5.1.1 Construction of the city-region specific Google Trends series 77

3.6 Results 81

3.6.1 Sentiment comparison 81

3.6.2 Test for Stationarity 82

3.6.3 Evaluation of the sentiment impact 83

3.6.4 Forecast 87

3.6.5 Robustness checks 102

3.6.5.1 Sentiment comparison: Macroeconomic indicator 102

3.6.5.2 Sentiment comparison: Office indicator 106

3.6.5.3 Sentiment comparison: Property specific indicators 107

3.6.5.4 Slicing 110

3.7 Conclusion 116

4 NATURAL LANGUAGE PROCESSING 119

4.1 Introduction 119

4.2 Literature review: Textual sentiment analysis 121

4.2.1 Natural Language Processing: Background 122

4.2.2 Sentiment analysis 124

4.2.3 NLP on the real estate market 131

4.2.4 NLP: Methodological development 133

4.3 Theory 135

4.4 Data description 137

4.5 Empirical framework 143

4.5.1 Base Model 143

4.5.2 Terminology 145

4.5.2.1 Corpus 145

4.5.2.2 Tokenization 145

4.5.2.3 Normalization and stemming 146

4.5.2.4 Lemma 147

4.5.3 Pre-Processing: Example 147

4.5.4 Sentiment extraction 148

4.5.4.1 AFINN 149

4.5.4.2 BING 150

4.5.4.3 NRC 150

4.5.4.4 Topic Modelling (*TM*) 151

4.6 Results 152

4.6.1 Autoregressive model 152

4.6.2 Robustness check 159

4.7 Conclusion 160

5 MACHINE LEARNING APPLICATION 162

5.1 Introduction 162

5.2 Literature review 165

5.3 Data Description 167

5.3.1 News Articles: Test dataset 167

5.3.2 Amazon data: Training dataset 169

5.3.3 Financial Times data 172

5.3.4 *MSCI* data 172

5.4 Empirical framework 174

5.4.1 Algorithms 175

5.4.2 Probit Models 176

5.5 Theoretical expectations 178

5.6 Results 179

5.6.1 Application of Amazon Book Reviews 179

5.6.1.1 Performance analysis 180

5.6.1.1.1 Training Data: Performance analysis 180

5.6.1.2 Graphical Analysis 187

5.6.1.2.1 All articles 188

5.6.1.2.2 No Housing articles 198

5.6.1.2.3 London 207

5.6.1.2.4 Newspapers with a circulation above 100,000 issues 215

5.6.1.2.5 Financial Times 222

5.6.1.2.6 Summary 231

5.6.1.3 Correlation analysis between the RICS U.K. commercial market survey and the textual sentiment indicators 233

5.6.1.4 Probit model 236

5.6.1.4.1 Sub-corpus I: All articles 236

5.6.1.4.2 Sub-Corpus II: No housing 255

5.6.1.4.3 Sub-Corpus III: London 270

5.6.1.4.4 Sub-Corpus IV: Newspapers with a circulation above 100,000 285

5.6.1.4.5 Sub-Corpus V: Financial Times 300
5.6.1.5 Robustness checks 315
5.6.1.5.1 Robustness check 1: Application of the textual sentiment indicators to more London specific series 317
5.6.1.5.2 Robustness check 2: Comparison between the RICS survey measures and the supervised learning measures in a probit model 322
5.6.1.5.3 Robustness Check 3: Comparison to the macroeconomic sentiment indicators and textual sentiment indicators from the previous parts 324
5.6.2 Development of a different training dataset using the lexicon approach 327
5.6.2.1 Performance analysis 331
5.6.2.2 Graphical interpretation 337
5.6.2.3 Fleiss and Cohen's Kappa 344
5.6.2.4 Implication into the probit model 346
5.7 Conclusion 349
6 CONCLUSION 353
6.1 An overview of the thesis 353
6.2 Limitations and future work 363
7 REFERENCES 367
8 APPENDIX i
8.1.1 Algorithms xxiv
8.1.1.1 Support Vector Machine (*SVM*) xxiv
8.1.1.2 Maximum Entropy Classifier (*MAXENT*) xxxiv
8.1.1.3 Stabilized Linear Discriminant Analysis (SLDA) xxxvi
8.1.1.4 Lasso and Elastic-Net Generalized Linear Models (*GLMENT*) xl
8.1.1.5 Decision *TREE* xlii
8.1.1.6 BOOSTING xlvi
8.1.1.7 BAGGING: Bootstrap Aggregation li
8.1.1.8 RANDOM FOREST lv
8.1.1.9 Neural Networks (*NNET*) lix

LIST OF TABLES

Table 3:1 - List of all countries and city-regions 39
Table 3:2 - Correlation (macroeconomic sentiment) 55
Table 3:3 - Regression results of the orthogonalization process (macroeconomic sentiment) 56
Table 3:4 - Principal component analysis (macroeconomic sentiment) 58
Table 3:5 - Correlation between the residuals and the first component 60
Table 3:6 - Correlation between the IPD total return index and the six office factors 62
Table 3:7 - Orthogonalization process (office sentiment) 63
Table 3:8 - Orthogonalization process (retail sentiment) 65
Table 3:9 - Summary of statistics 69
Table 3:10 - Correlation analysis 81
Table 3:11 - Fisher's Unit root test 83
Table 3:12 – Panel regression results: office yield model 84
Table 3:13 - Panel regression results: retail yield model 86
Table 3:14 - Forecast evaluation (office models) 88
Table 3:15 - Regional forecast evaluation: office, base model I 89
Table 3:16 - Regional forecast evaluation: office, base model II 90
Table 3:17 - Regional forecast evaluation: office, ME sentiment model I 91
Table 3:18 - Regional forecast evaluation: office, ME sentiment model II 92
Table 3:19 - Regional forecast evaluation: office, office sentiment model 93
Table 3:20 - Regional forecast evaluation: office, Google Trends I 94
Table 3:21 - Regional forecast evaluation: office, Google Trends II 95
Table 3:22 - Forecast evaluation (retail model) 96
Table 3:23 - Regional forecast evaluation: retail, base model 98
Table 3:24 - Regional forecast evaluation: retail, ME sentiment 99
Table 3:25 - Regional forecast evaluation: retail, retail sentiment 100
Table 3:26 - Regional forecast evaluation: retail, Google Trends 101
Table 3:27 - Robustness check: ME sentiment comparison, office yield 104
Table 3:28 - Robustness check: ME sentiment comparison, retail yield 105
Table 3:29 - Robustness check: office sentiment, office yield 106
Table 3:30 - Correlation analysis 107
Table 3:31 - Robustness check: property sentiment, office yield 108
Table 3:32 - Robustness check: property sentiment, retail yield 109
Table 3:33 - Robustness checks: slicing (GUF), 0ffice yield model 111
Table 3:34 - Robustness checks: slicing (GUF), retail yield model 112
Table 3:35 - Robustness checks: slicing (rEUR), office yield model 113
Table 3:36 - Robustness checks: slicing (rEUR), retail yield model 114

Table 4:1 - Overview of all collected market reports *138*
Table 4:2 - Overview of the planned analysis *140*
Table 4:3 - Summary of statistics: NLP *141*
Table 4:4 - Augmented Dickey-Fuller Test *142*
Table 4:5 - Overview of the different lexicons *152*
Table 4:6 - Result for the AR (1) model: overall commercial document corpus *155*
Table 4:7 - Result for the AR (1) model: all office related market reports *156*
Table 4:8 - Result for the AR (1) model: all office related market reports for London *158*
Table 4:9 - Robustness check: correlation analysis (RICS) *159*
Table 5:1 - Amazon book review training corpus *171*
Table 5:2 - Transformation of the categories *171*
Table 5:3 - Example of the range of rantings *172*
Table 5:4 - Descriptive statistics for the dependent variable *173*
Table 5:5 - Performance analysis: five classes *185*
Table 5:6 - Performance analysis: three classes *186*
Table 5:7 - Correlation analysis - lexicon approach - (all articles) *189*
Table 5:8 - Correlation analysis - supervised learning approach - (all articles) - 5 categories - all reviews *191*
Table 5:9 - Correlation analysis - supervised learning approach - (all articles) - 5 categories - equal number of reviews *193*
Table 5:10 - Correlation analysis - supervised learning approach - (all articles) - 3 categories - all reviews *195*
Table 5:11 - Correlation analysis - supervised learning approach - (all articles) - 3 categories - equal number of reviews *197*
Table 5:12 - Correlation analysis - lexicon approach - (no housing) *199*
Table 5:13 - Correlation analysis - supervised learning approach - (no housing) - 5 categories - all reviews *201*
Table 5:14 - Correlation analysis - supervised learning approach - (no housing) - 5 categories - equal number of reviews *203*
Table 5:15 - Correlation analysis - supervised learning approach - (no housing) - 3 categories - all reviews *204*
Table 5:16 - Correlation analysis - supervised learning approach - (no housing) - 3 categories - equal number of reviews *206*
Table 5:17 - Correlation analysis - lexicon approach - (London) *208*
Table 5:18 - Correlation analysis - supervised learning approach - (London) - 5 categories - all reviews *209*
Table 5:19 - Correlation analysis - supervised learning approach (London) - 5 categories equal - equal number of reviews *211*
Table 5:20 - Correlation analysis supervised learning approach - (London) - 3 categories - all reviews.. *214*

Table 5:21 - Correlation analysis -supervised learning approach - (London) - 3 categories - equal number of reviews *214*
Table 5:22 - Correlation analysis - lexical indicators - (100,000) *216*
Table 5:23 - Correlation analysis - supervised learning approach - (100,000) - 5 categories - all reviews *217*
Table 5:24 - Correlation analysis - supervised learning approach - (100,000) - 5 categories - equal number of reviews *219*
Table 5:25 - Correlation analysis -supervised learning approach - (100,000) - 3 categories - all reviews *220*
Table 5:26 - Correlation analysis - supervised learning approach - (100,000) - 3 categories - equal number of reviews *221*
Table 5:27 - Correlation analysis among the lexical indicators (FT) *223*
Table 5:28 - Correlation analysis - supervised learning approach - (FT) - 5 categories - all reviews *225*
Table 5:29 - Correlation analysis -supervised learning approach - (FT) - 5 categories - equal number of reviews *226*
Table 5:30 - Correlation analysis - supervised learning approach - (FT) - 3 categories - all reviews *228*
Table 5:31 - Correlation analysis - supervised learning approach - (FT) - 3 categories - equal number of reviews *230*
Table 5:32 - Correlation between leading indicators *232*
Table 5:33 - Correlation table between the AFINN, BING and MAXENT I indicators and the U.K. RICS survey measures *235*
Table 5:34 - Summary of statistics (all articles) *237*
Table 5:35 - Augmented Dickey-Fuller Test (all articles) *238*
Table 5:36 - Probit results: MSCI - all assets - all properties (all articles) *239*
Table 5:37 - Probit results: MSCI - all assets - all offices (all articles) *242*
Table 5:38 - Diebold-Mariano Test - MSCI all properties all assets (all articles) *250*
Table 5:39 - Diebold Mariano Test - MSCI all properties all offices (all articles) *250*
Table 5:40 - Forecast evaluation for the three turning points of the MSCI all properties series (all articles) *251*
Table 5:41 - Forecast evaluation for the three turning points MSCI all offices (all articles) *253*
Table 5:42 - Summary of statistics (no housing) *255*
Table 5:43 - Augmented Dickey-Fuller Test (no housing) *256*
Table 5:44 - Probit results: MSCI - all assets - all properties (no housing) *258*
Table 5:45 - Probit results: MSCI - all assets - all office properties (no housing) *260*
Table 5:46 - Diebold Mariano Test - MSCI all properties (no housing) *265*
Table 5:47 - Diebold Mariano Test - MSCI all offices (no housing) *265*
Table 5:48 - Forecast evaluation for the three turning points MSCI all properties (no housing) *266*
Table 5:49 - Forecast evaluation for the three turning points MSCI all offices (no housing) *268*
Table 5:50 - Summary of statistics (London) *270*

Table 5:51 - Augmented Dickey-Fuller Test (London) *271*
Table 5:52 - Probit results: MSCI - all assets - all properties (London) *273*
Table 5:53 - Probit results MSCI - all assets - all office properties (London) *275*
Table 5:54 - Diebold Mariano Test - MSCI all properties (London) *280*
Table 5:55 - Diebold Mariano Test - MSCI all offices (London) *280*
Table 5:56 - Forecast evaluation for the three turning points - MSCI all properties (London) *281*
Table 5:57 - Forecast evaluation for the three turning points - MSCI all offices (London) *283*
Table 5:58 - Summary of statistics (100,000) *285*
Table 5:59 - Augmented Dickey-Fuller Test (100,000) *286*
Table 5:60 - Probit results: MSCI - all assets - all properties (100,000) *288*
Table 5:61 - Probit results: MSCI - all assets - all offices (100,000) *290*
Table 5:62 - Diebold Mariano Test - MSCI all properties (100,000) *295*
Table 5:63 - Diebold Mariano Test - MSCI all offices (100,000) *295*
Table 5:64 - Forecast evaluation for the three turning points - MSCI all properties (100,000) *296*
Table 5:65 - Forecast evaluation for the three turning points - MSCI all offices (100,000) *298*
Table 5:66 - Summary of statistics (FT) *300*
Table 5:67 - Augmented Dickey-Fuller Test (FT) *301*
Table 5:68 - Probit results: MSCI - all assets all properties (FT) *303*
Table 5:69 - Probit results: MSCI - all assets - all office properties (FT) *305*
Table 5:70 - Diebold Mariano Test - MSCI all properties (FT) *310*
Table 5:71 - Diebold Mariano Test - MSCI all offices (FT) *310*
Table 5:72 - Forecast evaluation for the three turning points - MSCI all properties (FT) *311*
Table 5:73 - Forecast evaluation for the three turning points - MSCI all offices (FT) *313*
Table 5:74 - Comparison of the regression results for the AFINN, BING and MAXENT I models *318*
Table 5:75 - Probit model RICS vs best indicators *323*
Table 5:76 - Robustness check 3 - sentiment indicators within a standard yield model *325*
Table 5:77 - Performance analysis – FT news corpus annotated with the sentiment lexicons *333*
Table 5:78 - Overall performance comparison between the Amazon book review and the lexical approach *334*
Table 5:79 - Performance analysis of the FT test dataset *336*
Table 5:80 - Correlation analysis - between new classifiers and labels from the lexicon approach *342*
Table 5:81 – Correlation analysis - between the new and the original classifiers *343*
Table 5:82 - Interpretation of Fleiss Kappa *344*
Table 5:83 - Interpretation of Cohen's kappa *345*
Table 5:84 - Fleiss kappa for newly constructed classifiers - including all classifiers *345*
Table 5:85 - Fleiss kappa for newly constructed classifiers - without the poor performer *346*
Table 5:86 – Cohen's Kappa for newly constructed classifiers and the basic lexicon classification *346*
Table 5:87 - Probit regression results for the newly constructed supervised learning algorithms *348*

Table 8:1 - Scoring coefficients (macroeconomic sentiment - Kaiser Criterion) ..... *i*
Table 8:2 - Correlation between the various residuals and the components (macroeconomic sentiment - Kaiser Criterion) ..... *i*
Table 8:3 - Correlation analysis (macroeconomic sentiment - Kaiser Criterion) ..... *ii*
Table 8:4 - Calculated weight for final sentiment construction (macroeconomic sentiment - Kaiser Criterion) ..... *ii*
Table 8:5 - PCA of the sentiment proxies (macroeconomic sentiment - PCA) ..... *ii*
Table 8:6 - Scoring coefficients (macroeconomic sentiment - PCA) ..... *ii*
Table 8:7 - Orthogonalization process (office sentiment II) ..... *iii*
Table 8:8 - PCA of the sentiment proxies (property sentiment I) ..... *iii*
Table 8:9 - Scoring coefficients for all components (property sentiment I) ..... *iii*
Table 8:10 - Correlation analysis (property sentiment I) ..... *iv*
Table 8:11 - Variable definition for the yield models ..... *iv*
Table 8:12 - Data description ..... *v*
Table 8:13 - Descriptive statistics (1) ..... *vi*
Table 8:14 - Descriptive statistics (2) ..... *vii*
Table 8:15 - Google Trends indicator construction ..... *viii*
Table 8:16 - Google Trends results for each city region ..... *ix*
Table 8:17 - Regional fixed effects for the office yield model (1) ..... *x*
Table 8:18 - Regional fixed effects for the office yield model (2) ..... *xi*
Table 8:19 - Regional fixed effects for the office yield model (3) ..... *xii*
Table 8:20 - Regional fixed effects for the office yield model (4) ..... *xiii*
Table 8:21 - Regional fixed effects for the retail yield model (1) ..... *xiv*
Table 8:22 - Regional fixed effects for the retail yield model (2) ..... *xv*
Table 8:23 - - Regional fixed effects for the retail yield model (3) ..... *xvi*
Table 8:24 - Regional fixed effects: office yield model (GERUKFRA) (I) ..... *xvii*
Table 8:25 - Regional fixed effects: office yield model (GERUKFRA) (II) ..... *xviii*
Table 8:26 - Regional fixed effects: retail yield model (GERUKFRA) ..... *xix*
Table 8:27 - Regional fixed effects: office yield model (rEUR) (I) ..... *xx*
Table 8:28 - Regional fixed effects: office yield model (rEUR) (II) ..... *xxi*
Table 8:29 - Regional fixed effects: retail yield model (rEUR) (I) ..... *xxii*
Table 8:30 - Regional fixed effects: retail yield model (rEUR) (II) ..... *xxiii*
Table 8:31 - Robustness check I (all) ..... *lxiii*
Table 8:32 - Robustness Check 1 (no housing) ..... *lxiv*
Table 8:33 - Robustness Check 1 (London) ..... *lxv*
Table 8:34 - Robustness Check 1 (100,000) ..... *lxvi*
Table 8:35 - Robustness Check 1 (FT) ..... *lxvii*

LIST OF FIGURES

Figure 1:1 - Survey based sentiment indicator (no event) 3
Figure 1:2 - Survey based sentiment indicator (event) 4
Figure 1:3 - Sentiment influenced by an event and the news coverage 5
Figure 3:1 - Gram-Schmidt Algorithm 54
Figure 3:2 - Orthogonalization process 57
Figure 3:3 - Scree plot of eigenvalues after PCA (macroeconomic sentiment) 59
Figure 3:4 - Orthogonalization process: IPD total return index (offices) for Berlin 64
Figure 3:5 - Orthogonalization process: IPD total return index (retail) for London West End 66
Figure 3:6 - Sentiment comparison for the London West End market 67
Figure 3:7 - Google Trends - "office" 73
Figure 3:8 - Google Trends - "Büro" 74
Figure 3:9 - Google Trends - "Büro" vs. "office" 74
Figure 3:10 - Google Trends - Regional interest 75
Figure 3:11 - Google Trends - City list 75
Figure 3:12 - Global market share of desktop search engines 77
Figure 4:1 - Number of market reports per year 139
Figure 4:2 - AFINN example 149
Figure 4:3 - BING example 150
Figure 4:4 - NRC example 151
Figure 4:5 - Topic modelling example 151
Figure 5:1 - Number of articles per sub-corpora per quarter 169
Figure 5:2 - Rating of the reviews 170
Figure 5:3 - Graphical illustration of the supervised learning approach 175
Figure 5:4 - Graphical illustration of precision and recall 182
Figure 5:5 - Lexicon approach (all articles) 189
Figure 5:6 - Classifiers trained on all book reviews: five classes (all articles) 190
Figure 5:7 - Classifiers trained on an equal number of book reviews: five classes (all articles) 192
Figure 5:8 - Classifiers trained on all book reviews: three classes (all articles) 194
Figure 5:9 - Classifiers trained on an equal number of book reviews: three classes (all articles) 196
Figure 5:10 - Lexicon approach (no housing) 198
Figure 5:11 - Classifiers trained on all book reviews: five classes (no housing) 200
Figure 5:12 - Classifiers trained on an equal number of book reviews: five classes (no housing) 202
Figure 5:13 - Classifiers trained on all book reviews: three classes (no housing) 204
Figure 5:14 - Classifiers trained on an equal number of book reviews: three classes (no housing) 205
Figure 5:15 - Lexicon approach (London) 207
Figure 5:16 - Classifiers trained on all book reviews: five classes (London) 208

Figure 5:17 - Classifiers trained on an equal number of book reviews: five classes (London) 210
Figure 5:18 - Classifiers trained on all book reviews: three classes (London) 212
Figure 5:19 - Classifiers trained on an equal number of book reviews: three classes (London) 213
Figure 5:20 - Lexicon approach (100,000) 216
Figure 5:21 - Classifiers trained on all book reviews: five classes (100,000) 217
Figure 5:22 - Classifiers trained on an equal number of book reviews: five classes (100,000) 218
Figure 5:23 - Classifiers trained on all book reviews: three classes (100,000) 220
Figure 5:24 - Classifiers trained on an equal number of book reviews: three classes (100,000) 221
Figure 5:25 - Lexicon approach (FT) 223
Figure 5:26 - Classifiers trained on all book reviews: five classes (FT) 224
Figure 5:27 - Classifiers trained on an equal number of book reviews: five classes (FT) 225
Figure 5:28 - Classifiers trained on all book reviews: three classes (FT) 227
Figure 5:29 - Classifiers trained on an equal number of book reviews: three classes (FT) 229
Figure 5:30 - Prediction of the MSCI all properties series - lexicon approach (all articles) 245
Figure 5:31 - Prediction of the MSCI all properties series - machine learning approach (all articles) 246
Figure 5:32 - Prediction of the MSCI all offices series - lexicon approach (all articles) 247
Figure 5:33 - Predictions of the MSCI all offices series - machine learning approach (all articles) 248
Figure 5:34 - Turning point predictions MSCI all properties (all articles) 252
Figure 5:35 - Turning point predictions MSCI all offices (all articles) 254
Figure 5:36 - Predictions of the MSCI all properties indicator - lexicon approach (no housing) 261
Figure 5:37 - Predictions of the MSCI all properties indicator - machine learning approach (no housing) 262
Figure 5:38 - Predictions of the MSCI all offices indicator - lexicon approach (no housing) 263
Figure 5:39 - Predictions of the MSCI all offices indicator - machine learning approach (no housing) 264
Figure 5:40 - Turning point predictions MSCI all properties (no housing) 267
Figure 5:41 - Turning point predictions MSCI all offices (no housing) 269
Figure 5:42 - Predictions of the MSCI all properties indicator - lexicon approach (London) 276
Figure 5:43 - Predictions of the MSCI all properties indicator - machine learning approach (London) 277
Figure 5:44 - Predictions of the MSCI all offices indicator - lexicon approach (London) 278
Figure 5:45 - Predictions of the MSCI all offices indicator - machine learning approach (London) 279
Figure 5:46 - Turning point predictions, MSCI all properties (London) 282
Figure 5:47 - Turning point predictions, MSCI all offices (London) 284
Figure 5:48 - Predictions of the MSCI all properties indicator - Lexicon approach (100,000) 291
Figure 5:49 - Predictions of the MSCI all properties indicator - Machine learning approach (100,000) 292
Figure 5:50 - Predictions of the MSCI all offices indicator - Lexicon approach (100,000) 293
Figure 5:51 - Predictions of the MSCI all offices indicator - Machine learning approach (100,000) 294
Figure 5:52 - Turning point predictions, MSCI all properties (100,000) 297
Figure 5:53 - Turning point predictions, MSCI all offices (100,000) 299

Figure 5:54 - Predictions of the MSCI all properties indicator - lexicon approach (FT) *306*
Figure 5:55 - Predictions of the MSCI all properties indicator - machine learning approach (FT) *307*
Figure 5:56 - Predictions of the MSCI all offices indicator - lexicon approach (FT) *308*
Figure 5:57 - Predictions of the MSCI all offices indicator - machine learning approach (FT) *309*
Figure 5:58 - Turning point predictions, MSCI all properties (FT) *312*
Figure 5:59 - Turning point predictions, MSCI all offices (FT) *314*
Figure 5:60 - Robustness Check I - BING model – pseudo-R-squared value comparison *319*
Figure 5:61 - Robustness Check I - AFINN model - pseudo-R-squared value comparison *320*
Figure 5:62 - Robustness Check I - MAXENT I model - pseudo-R-square value comparison *321*
Figure 5:63 - New FT training corpus *329*
Figure 5:64 - Distribution of the FT corpus over the three different classes *330*
Figure 5:65 - NRC - Classifiers trained on an FT news corpus *338*
Figure 5:66 - TM - Classifiers trained on an FT news corpus *339*
Figure 5:67 - AFINN - Classifiers trained on an FT news corpus *340*
Figure 5:68 - BING - Classifiers trained on an FT news corpus *341*
Figure 8:1 - Geometric interpretation of standard SVM ..... *xxv*
Figure 8:2 - Non-linear separable data ..... *xxix*
Figure 8:3 - Kernel function applied ..... *xxx*
Figure 8:4 - One-versus-all approach ..... *xxxii*
Figure 8:5 - Application of the Fisher LDA ..... *xxxvii*
Figure 8:6 - Example of the different penalties ..... *xli*
Figure 8:7 - Structure of a decision TREE ..... *xliii*
Figure 8:8 - Classification categories based on their error rate ..... *xlvi*
Figure 8:9 - Simple neural network consisting of two neurons ..... *lix*

1 INTRODUCTION

1.1 BACKGROUND AND MOTIVATION

The Efficient Market Hypothesis of Fama (1970) states that asset prices reflect all available market information and only change when new information enters the market. This hypothesis, as well as other classic financial theories, such as the Capital Asset Pricing Model or the Arbitrage Pricing Theory, have dominated the finance world, and alternative theories have struggled to be accepted in academia. Such theories require the belief that market participants base their decisions on a rational framework and act as rational and return-maximizing investors. Due to the difficulties in explaining certain recurring phenomena, such as the January Effect or the Equity Premium Puzzle, which do not fit into this framework, researchers tried to develop an alternative approach. A number of studies have revealed that rationality within the market is less present than assumed and that static models can be improved when more realistic assumptions, such as the so-called human element, are considered. Behavioural finance has been developed over a long time and included psychological elements to justify the specific irrational behaviour of investors. The field has changed the focus towards the individual and his or her actions within the market. Especially in the last decade, new research methods and new datasets have helped to develop the field and have been put on the research agenda.

One measure of the so-called human element is market sentiment. According to Baker and Wurgler (2007), sentiment is the belief of investors about future cash flows and the investment risk that is not justified by the facts at hand. In other words, sentiment describes the belief about future developments of the market. This is based on all collected information and how it is processed and rated within the mind of the individual.

The literature differentiates between two groups of sentiment measures. The first group uses interviews and surveys to extract the beliefs from market participants. Since the measure is built on the direct interaction with market participants, direct sentiment indicators provide the best indication of future developments. However, these surveys require constant maintenance and the willingness of the interviewees to take part in the process. The construction of survey-based measures can also be described as time-consuming. Another issue which arises when direct sentiment measures are used in multinational studies is the fact that direct measures are not always comparable to each other. The main reason can be the difference in the underlying structure of the questionnaire. Prominent examples of direct

measures are the Economic Sentiment Indicator (ESI) [Tsolacos (2012)], the published sentiment surveys of the RICS, the survey of the Real Estate Research Corporation (RERC) [Clayton, Ling and Naranjo (2009); Freybote (2016)], the Conference Board Consumer Confidence Index [Bram and Ludvigson (1997); Howrey (2001)] and the University of Michigan Consumer Sentiment Index, first published by Katona (1947) and later used by Carroll et al. (1994) and Marcato and Nanda (2016).

The second group of sentiment measures utilizes the fact that direct measures are not always available. A variety of studies have used indirect sentiment indicators to measure the underlying market sentiment [Choi and Varian (2009), Preis et al. (2010), Freybote and Seagraves (2017), Baker and Wurgler (2006)]. However, indirect sentiment indicators do not measure the sentiment in the first place. With different statistical methods, the assumed sentiment is extracted from these proxies (i.e. orthogonalization). Unfortunately, it remains questionable whether an orthogonalized sentiment indicator actually measures the sentiment. For instance, Clayton et al. (2009) compared a sentiment proxy to the RERC survey and found contradicting results. The main problem when conventional sentiment proxies are used is the time difference between the measured sentiment and the publication date of the indicators. In order to generate the indicators, the proxy measures have to be published first. This generates a time lag, and uncertainty about the market arises.

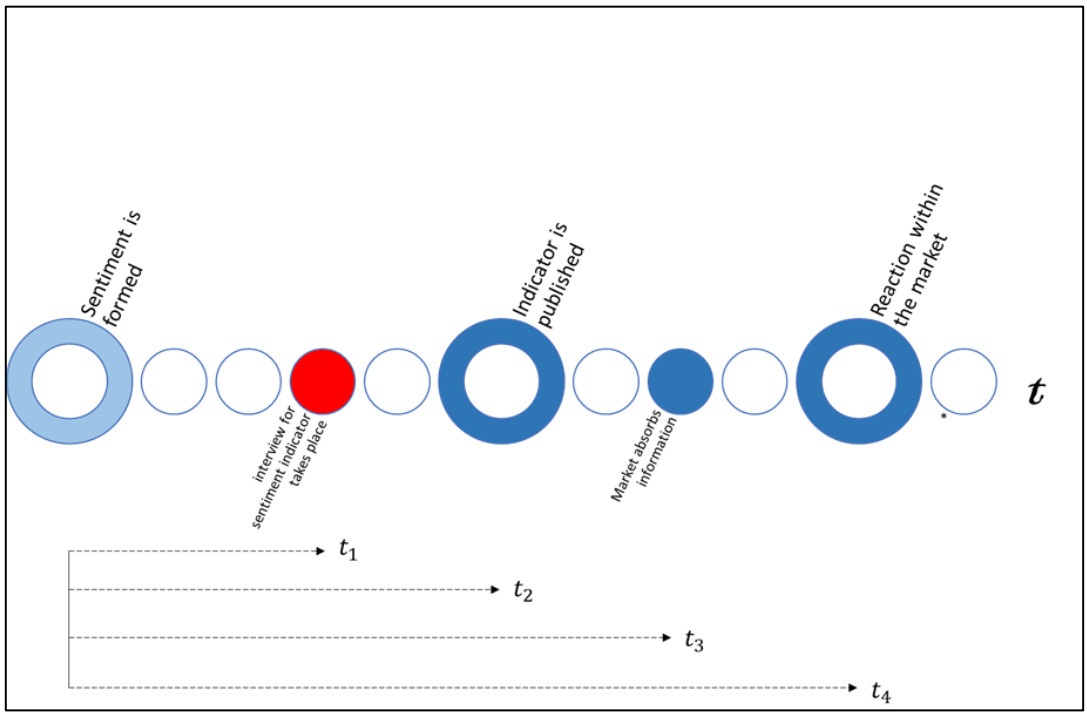
The literature shows that surveys provide a better market sentiment than indirect measures. However, they should also be treated with caution. The group of interviewees influences the outcome of the survey tremendously. I further see the time gap between the data collection and the publication of the results as a possible window of misinformation and noise.

The following two figures illustrate the different time periods involved in the process of sentiment extraction. Two layers are essential, the personal layer of the interviewee and the market layer where the aggregated sentiment is absorbed. It is assumed that multiple individuals share a common sentiment and that the sentiment indicator will reflect the aggregated opinion of the market.

After the indicator is published, it is further assumed that market participants absorb this published opinion and change their behaviour accordingly. It is also presupposed that, between the interview and the publication of the indicator, no significant event has taken place (Figure 1:1). In the case of a new event (Figure 1:2), the sentiment would have been different from that

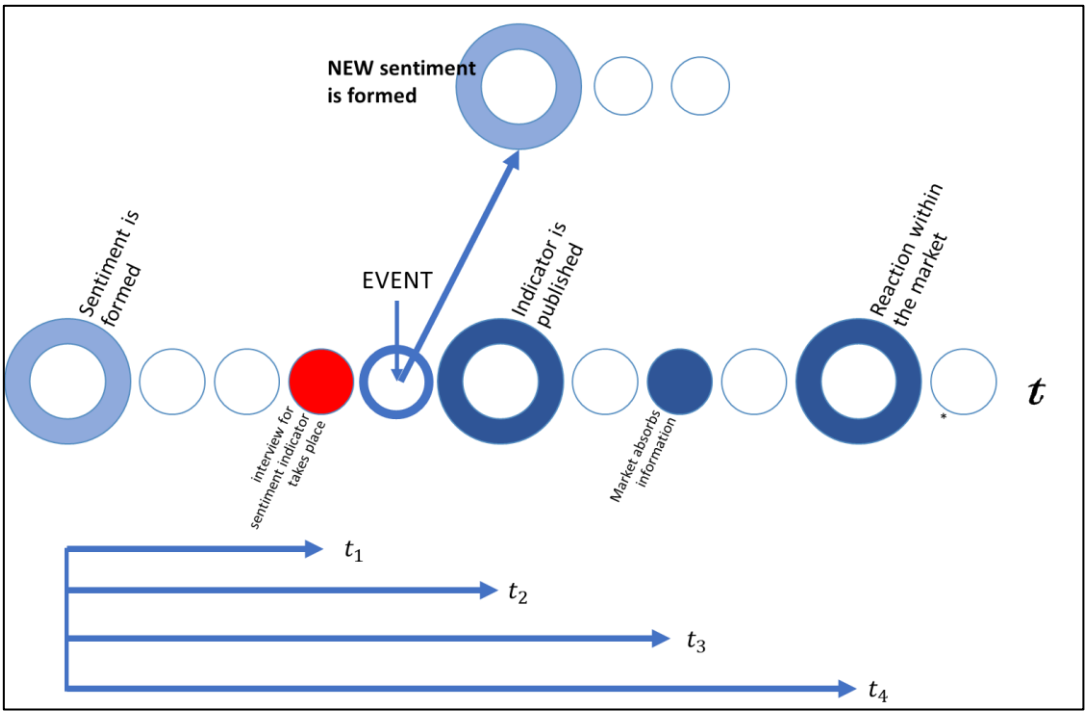
moment onward, and the published indicator provides a wrong or outdated signal to the market.

Figure 1:1 - Survey based sentiment indicator (no event)



Note 1.1: The figure illustrates an Idealised process of a sentiment extraction with the help of a survey. It is assumed, that the sentiment, which has been formed by the individual interviewees before the interview, is multiplied by the publication of the survey results. The market will absorb and react to the assumed "market sentiment".

Figure 1:2 - Survey based sentiment indicator (event)



Note 1.2: Different to Figure 1:1 the idealised process is disturbed by an unexpected event, which takes place between the interview and the publication. Therefore, the results of the survey will report an outdated market sentiment.

Since the literature has not come up with a universal sentiment proxy, which could be applied to different markets, in this thesis I try to supply an updated approach for the use of sentiment proxies. I have identified three areas which contribute to the decision making of market professionals. I assume that they either (1) consult friends or colleagues, (2) rely on their experience or (3) that they consume various information to make a sound decision. Since the first two points are difficult to measure in a scientific framework, I will rely on sentiment extracted from text documents.

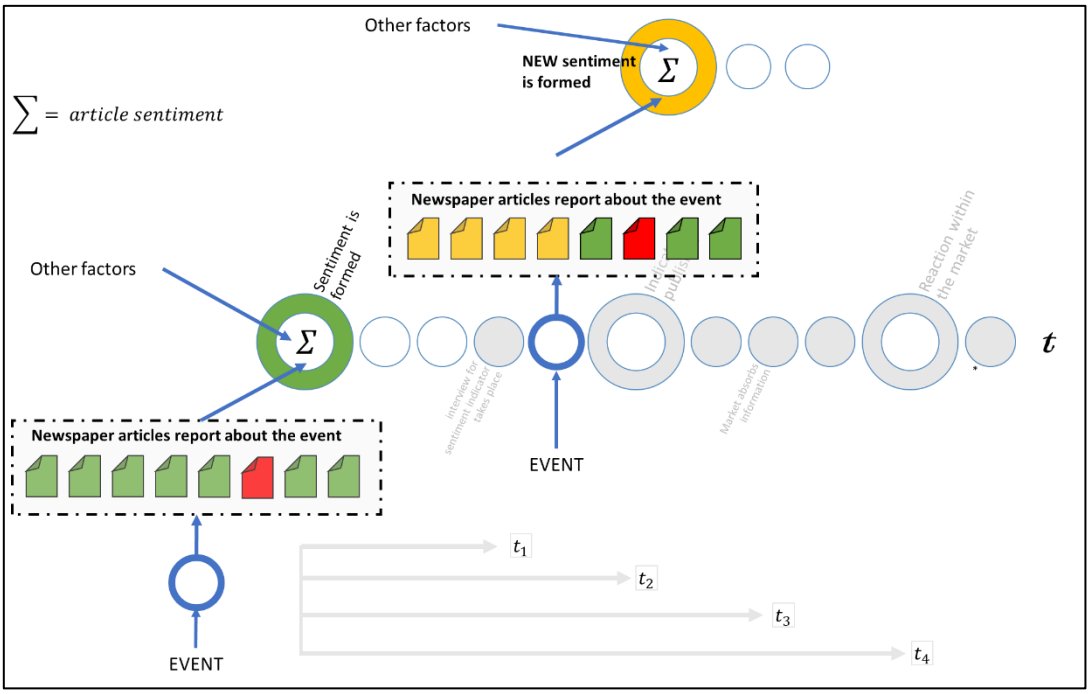
Text documents have the advantage of reflecting the market and its developments much closer to a specific moment in time. However, sentiment extracted from texts does reflect the opinion of an individual author who describes the current market situation and, in some cases, provides an *ex-ante* indication. Macroeconomic sentiment indicators are based on proxies which are measured *ex-post*. In this thesis, I will use market reports both from service agencies and newspaper articles.

Journalists of the latter category try to give an objective description of an event or topic. However, they are also driven by other aspects, which influence their writing style and the message they provide. Besides an informative function, they also have to entertain and make

sure that readers are attracted and bonded to the newspaper. The developed textual sentiment indicators are in general based on the wording of these articles.

Figure 1:3 illustrates on which base the sentiment is likely to be influenced. As stated above, the idealised process could be disturbed by an event. This event might shift the sentiment from several market participants. The reported sentiment index based on the survey could therefore be outdated. Newspaper articles or other text documents report on the development of the market constantly. If a market participant is reading a range of articles concerning the event, he might change his opinion and sentiment about the market development, based on the underlying sentiment in the articles. However, as stated before, the sentiment is also influenced by other factors as well.

Figure 1:3 - Sentiment influenced by an event and the news coverage



Note 1.3: The figure is based on the original process of survey extraction via a survey (Figure 1:1). As shown in Figure 1:2 this process is disturbed or ends in an outdated sentiment measure. The above-presented figure, is added by a possible source, which influences the sentiment of the market participants. News articles, or text documents in general, will report on these events. As presented the aggregated view of the documents (colour) will among other factors, influence the newly formed sentiment.

1.2 AIMS AND OBJECTIVES

The aim of this thesis is to analyse and measure the sentiment on the European commercial real estate market. It is my opinion, that policymakers and market participants could benefit from a deeper insight in the market sentiment. While, sentiment can be measured in different ways, it is essential to realise that each method has its advantages and disadvantages. In this thesis I tackle two issues. First, while direct sentiment measures are costly and time-consuming to construct, they are seldom available for multiple regions or even countries. This prohibits a comparison between different markets. Second, as I have just shown, those measures are likely to be out-dated, when they are published, since they refer to a sentiment, which has been formed before the interview took place. In this thesis I try to bridge those issues, by first establishing the need of a European wide sentiment measure and second offering a method which is able to provide an updated measure, which is much closer linked to the actual market development.

In more detail, the first part of the thesis tries to answer the questions if the European commercial real estate market is subject to sentiment? As there is no European wide real estate sentiment measure, I wonder, if a range of European wide sentiment proxies can provide an insight into the market of individual countries? Three objectives are pursued, first, the research attempts to show that sentiment extracted from a different set of proxies will provide sufficient information. Second, different methods will be tested in ordered to asses which method should be followed. In general, two approaches will be discussed. Depending on the specific sentiment measures either a principal component analysis or a two-stage method as a combination of orthogonalization and PCA will be tested. And third, due to data availability and complexity in the construction of the sentiment measures, a more straightforward approach, based on online search volume data, will be examined. The first part aims to establish the need of a generalized sentiment measure. Measures, which can be transferred from one market to another, allow market participants to draw more general conclusions and offer the possibility to compare different markets with each other. Direct sentiment measures, which are based on different time frames, target groups or question sets, do not offer these advantages.

Due to the heavy reliance on the availability of different sentiment proxies, the first part of the thesis will identify a time lag between, the sentiment of the market and the publication of the proxies. Driven by that, the second part of the thesis deals with the question, if there is an alternative which can provide a much more topical medium and method? Therefore, the thesis further tries to evaluate if the extraction of sentiment from text documents can provide a better

image about the development within the market? Here, the objective of the research is to associate new methods and data sources to the commercial real estate market. The second part, tries to utilize market reports from service agencies for the London commercial real estate market. The sentiment from these documents will be extracted by four different lexicon approaches. Their performance will be measured with the help of an autoregressive model. It is of interest to estimate which approach and which combination of reports provides a better market picture. The chapter does not try to provide a sufficient modelling framework, since the introduction of the method and the medium stands in the centre of interest. Market participants and policymakers will benefit from the consideration of text documents as a source of sentiment, since text documents are constantly published. Different to the first part, both the medium and the method, are much more straightforward, when it comes to modification and data handling.

Finally, the thesis tries to answer a series of different questions concerning newspaper articles and the application of supervised learning algorithms. The main goal of the third part of the thesis is it to answer the question if market participants change their behaviour based on the information they consume? In addition, do newspaper articles offer enough market noise, in order to extract sentiment from them? Newspaper articles are published with a higher frequency, in comparison to market reports. Therefore, sentiment extracted from those texts should be much closer to the actual market development. Besides the change of the text documents, the third part of the thesis introduces another method, which offers promising features for the extraction of sentiment. Nine different supervised learning methods will be applied in order to extract the sentiment from five different news corpora. One research objective is it to establish, what underlying focus the test dataset requires and which algorithm produces the best result. Five different sub-corpora have been constructed in order to answer this question. Besides these objectives, this part tries to provide an alternative approach when it comes to train the algorithms. Supervised learning algorithms require a training and a test dataset. Since, the real estate industry does yet not offer an adequate training dataset, I offer two alternatives to bridge this gap. The first question is, are Amazon real estate book reviews able to train supervised learning algorithms sufficiently? The second question is, can a combination of wordlists and supervised learning algorithms produce more robust results? Amazon book reviews are essentially classified texts, which can be used to train the different algorithms. While it might be a bit far-fetched, that the book reviews are similar to real estate related news articles in their wording, the second method utilises the wordlist approach to classify another set of news entities. This method has the advantage, that both the training and

test dataset are similar in style. The constructed sentiment measures and their performance will be tested in a probit framework.

Coming back to the essential questions of this thesis, I hope to provide enough knowledge to the field to allow different market players to utilise on my findings. Text documents, different to macroeconomic variables or sentiment surveys, are published in a constant manner in all countries. Therefore, the proposed methods should offer the advantage of transferability to other markets.

Before, I will describe the following chapters in more detail, I like to provide a short overview of the field of behavioural finance. Starting more general with the origins, I will point out, how behavioural finance has been applied to the field of real estate.

1.3 BEHAVIOURAL FINANCE ORIGINS

In 1952 the field of finance started to change completely. The late Nobel Prize winner Harry M. Markowitz published his idea of Modern Portfolio Selection (1952a), which adopts mathematical techniques to improve the investment process. The strategy of building diversified stock portfolios based on a mean-variance framework was further transformed in the following years. The Capital Asset Pricing Model (CAPM) was independently developed by Treynor (1962), Sharpe (1964), Lintner (1965) and Mossin (1966) and the Arbitrage Pricing Theory (APT) was presented by Ross in 1976. Market participants are assumed to be rational and risk averse at all times. The Efficient Market Hypothesis (EMH) of Eugene Fama (1970) allows for irrational investors. However, this group is needed to prove the theory right. Irrational investors are assumed to be the reason for prices in disequilibrium. They face rational arbitrageurs who will push prices back to equilibrium because of their superior knowledge.

Although these theories only work in an experimental environment, they have been used for many years with success. After traditional finance theories were established, alternative ones were not accepted for a long time. Behavioural interaction during the decision-making process was considered a possible explanation. Markowitz (1952b) for instance published

another paper which deals with the behaviour of people regarding their utility function. The paper tries to answer the question why some people buy insurances and show risk aversion, while others do not buy them and take riskier decisions. One reason, according to the author, is that people try to improve their wealth when they are unsatisfied with their current level. That could explain why some people have an irrational betting behaviour when they take bad bets with the possibility of more substantial returns. Markowitz based his idea on the work of Friedman and Savage (1948) who also discussed the choices under the influence of risk. The authors provided an in-depth behavioural analysis which led to further studies. They identified specific boundaries why some groups of society are not able to enter fair games and why other groups choose specific risks in their decisions. The reason for the latter can be the expected return, which increases at the same time as the risk increases. So, the attempt to explain the individual irrational behaviour of market participants was already present at the beginning of traditional finance. Other scholars who were motivated by those unrealistic assumptions or by the existence of market anomalies, which could not be explained by the traditional finance theories, tried to find a way to disprove those theories and to develop an alternative.

With the adoption of psychological and sociological points of view the field of behavioural finance evolved. The main advantage by adopting the views of those disciplines is the fact, that the basic assumptions of the traditional finance theories (e.g. the sole aim of investors to maximise their returns; or the ability to absorb and process all information immediately) are recognized as unrealistic. Due to the influence of psychological studies, researchers agreed on the fact that economic theories should put the individual and his or her behaviour at the centre of interest. Thaler (2010) argues that the “representative investor” is expected to be rational in a twofold way. On one hand, he bases his decisions on financial theories, and on the other, his predictions of the future are unbiased. Those unrealistic assumptions cannot hold in the real world. In Thalers (2010) opinion, behavioural finance has overcome the status of a controversial discipline and will replace the traditional theories. Unfortunately, Thaler does not explain which alternative theory investors should follow.

A precise definition of behavioural finance is hard to find since many scholars believe that the field is still in the fledgeling stages and changes continuously. This point of view can be confirmed if the variety of fields which are now contributing to behavioural finance are considered. Ricciardi and Simon (2000) gave an overview of research fields, including anchoring, information cascades, under-reaction and over-reaction, as well as risk perception. One attempt at a definition can be found in Park and Sohn (2013). Their exhaustive literature review identifies two stages of behavioural finance: a macro-stage and a micro-stage. Whereas the macro-stage

focuses on the observed anomalies in the Efficient Market Hypothesis which can be explained by behavioural finance, the micro-stage instead focuses on the individual and his or her biases towards specific behaviour. That shows that the field is concentrated on two aspects: the broader picture and the investor themselves. De Bondt et al. (2008) suggest that behavioural finance is based on three blocks: sentiment, behavioural preferences and limits to arbitrage. Another attempt with further detail can be found in Ricciardi and Simon (2000): they conclude that behavioural finance looks at the financial market from the perspective of an individual and tries to explain "the what, why and how of finance". Statman (1995), Barber and Odean (1999) and Shefrin (2000) also focus on the individual and how his or her decision-making process has been influenced by behaviour and psychology. They identify information processing and risk assessment as the primary drivers of behavioural finance. Investors should be aware of the human factor so as to avoid mistakes and to use the misjudgement of others to achieve an advantage, since misjudgement happens consistently. Fundamental work regarding the decision-making process was done by Simon (1957). His work on heuristics showed that the human brain tends to use only a subset of its potential to solve particular problems. To summarize, all authors agree that behavioural finance enriches our understanding of financial markets. What behavioural finance does not do is to give satisfying alternatives to the established models. It is therefore not clear how and when behavioural finance will replace the neoclassic approach as suggested by Thaler.

This is why established scholars such as Fama are still quite critical when it comes to the discussion. His critique in 1998 includes for instance that the observed over-reaction is balanced by the same amount of under-reaction. Furthermore, the discussion of the Efficient Market Hypothesis is often based on vague and short-term events which can be disproved in the long run. One principal argument of Fama (1998) is that such results are sensitive to the methodology which is used. He concludes, that against all the odds, such critics, including the field of behavioural finance, are unable to offer a better and generalized alternative, which is why the Efficient Market Hypothesis survives. The question stated at this point is, has behavioural finance ever claimed to develop an alternative regarding trading strategies or was the field developed to point out where standing theories show weaknesses to give an impulse for the improvement of those theories? One goal which has been achieved so far is the acceptance and incorporation of the human factor.

Ricciardi and Simon (2000) state that behavioural finance emerged in the early 1990s; however, given the evidence presented in this review, this assumption is wrong. As pointed out earlier scholars have worked on related topics since the early 1950s. Due to the dominance of

the traditional theories, research output was not as great as it has become during recent decades. Nevertheless, scholars such as Keynes, Knight, Markowitz, Friedman and Savage, as well as Popper and Katona have to be mentioned. They worked in the field of behaviour or at least in a related subject. Besides his achievements in the field of sentiment analysis, Katona (1953) also contributes to the discussion of how our behaviour influences our decisions. He assumes that behaviour is pre-programmed either by education or by inheritance. The fundamental principle of behaviour is, therefore, repetition and/or habits. So, it is not clear that traders ever could react rationally since they show biased behaviour in the first place. More recent studies have linked the behaviour of people to their genetic code. De Neve and Fowler (2014) highlighted that behaviour is to a certain extent predefined by the gene code.

Katona (1953) does not entirely agree with Markowitz (1952b) and Friedman and Savage (1948). For him, it is difficult to justify why certain individuals have a particular utility function, and it remains unclear why people tend to change it over time. In Hirshleifer and Shumway (2003) more evidence against the individual's specific utility function can be found. They proved that externalities such as the weather do have a strong influence on trading behaviour. They found that sunshine is strongly related to stock returns and could even further develop a trading strategy based on this relationship.

Shiller (2003) points out that early signs of more significant disagreement with the standard theories could be seen during the 1970s. Among others, Fama (1970) admitted the existence of anomalies, but argues that they are a necessary element of the Efficient Market Hypothesis. Shiller (2003) admits that smaller anomalies such as the *January Effect* or the *Day-of-the-week Effect* could be seen as marginal in proving the Efficient Market Hypothesis wrong, whereas the anomaly of excess volatility within returns cannot be neglected. Changes in prices occur without any primary backup and seem to follow "animal spirits". This phenomenon of noise traders was also discussed by De Long et al. (1990) and by Barber et al. (2009). In an Efficient Market Hypothesis framework, noise traders are assumed to be irrational and impulsive. Arbitrageurs are not able to react in the theoretical way because of the inherent risk of noise traders whose behaviour is impossible to predict. However, noise traders do provide themselves with more substantial returns in comparison to rational traders due to the risk they engage in. The reason for this imbalance lies in the fact that arbitrageurs focus on a short horizon and face liquidity problems in the long run. The above authors assume that rational traders are not only trading on fundamentals but invest more time in the analysis of noise trader behaviour so as to examine specific patterns. This behaviour can be assigned to chartists.

Conforming to the concluding remarks of Ricciardi and Simon (2000), representatives of both camps could agree on the relevant topics which should be taught in schools, where there is room for alternative theories such as the Prospect Theory of Kahneman and Tversky. Having their origin in the field of psychology, their research is still widely used as evidence against neoclassical finance assumptions. With the introduction of psychological techniques and experiments, the authors developed the Prospect Theory in 1979. The theory states that individuals use reference points before they decide; this explains how they evaluate choices with known risk probabilities for the outcomes. People tend to value the potential loss or gain more than the actual outcome. This observation confirms the general assumption that individuals may not react in an entirely rational way. The Prospect Theory was derived from their earlier work, the Theory of Subjective Probabilities (1972) and the Theory of Small Numbers (1971). Kahneman and Tversky (1972) state that people apply probabilities to an event by assuming that the probability can be transferred from the parent population. That leads to an incorrect decision since the size of a sample does not have any or at least only a small influence on the likelihood of an event. However, we tend to base the majority of our decisions on probabilities we have experienced or observed and sum them up. This is also called representativeness. In the theory of small numbers, the authors have proven that people tend to have a strong intuition about random sampling, which causes errors in the following, since their conclusions are based on a wrong sample size. Individuals believe that small samples drawn from a larger parent population are much more similar to the larger population than they are. This proves that decisions are based on non-rational assumptions which can be generalized to a variety of individuals. However, those results are based on experiments and should be treated with caution. Bosch-Domènech and Silvestre (2010), for instance, showed that findings that are based on experiments could lead to wrong results. For instance, the Prospect Theory has been proven wrong when the participants had to run the experiment with real money and had to face real losses. This showed that individuals are not at risk when facing high probability losses as was suggested by Kahneman and Tversky (1979). Posner (2012) also criticized the focus on the achievements of Kahneman. In his opinion, other scholars such as Shiller or Shleifer have contributed a more significant share to insight into the field. According to the author, they identified patterns where others assumed random behaviour, such as in the reluctance to sell loser stocks or the focus on “hot” stocks while ignoring long-run trends. The question at hand is, is herding behaviour – the selling of stocks when others sell, and the buying of stocks when others buy – irrational? One reason for following the herd can be found in our natural instincts. Another reason might be the logic that betting against the flow may cause more personal regret

when the trader is wrong at the end. It is easier to accept a mistake when you are part of the herd.

Another essential element of the neoclassic finance theories is that investors are assumed to collect all available information and that those hypothetical individuals are capable of processing an unlimited amount of information instantly. They are further assumed to be able and willing to update their information regularly and adjust their decisions. The first part of this hypothesis that people can process an unlimited amount of information was refuted by Miller (1956). He showed that the human brain could only process seven chunks of information at once. Many scholars, such as Rabin (1998), Camerer et al. (2003) or Shiller (2003), take up the position that information in the decision-making process hardly plays any role. According to Garcia (2013), individuals put more weight on information that is consistent with their preferences and either ignore or forget other information which is contradictory. Sometimes, individuals ignore all given information and base their decision on an impulse. Over-confidence can be seen as a primary driver of this.

Even if the assumption that information does not impact on the decision significantly was accepted, it goes without question that information does play a vital role in the investment process. Investors or individuals who face an investment decision at least try to be rational in the sense that all available information is gathered and analysed. It remains unclear how individuals process this information and whether it is used to adjust their behaviour. So far, this has been neglected in the literature.

The phenomenon of information cascades is observable and leads to irrational decisions. One reason for this can be found in Shiller et al. (1984) who describe investment as a social activity. Individuals talk about their successes and failures and exchange ideas about new possibilities. People tend to put more weight on the opinions of close friends or relatives; they also follow trends and fashions. The authors further point out that trends occur without any particular reason and move in some cases from one country to another. According to social scientists, one background mechanism for herding might be group pressure or the diffusion of opinions. Both lead to irrational reactions, whereas in the first case people do not like to be isolated or run against the flow, and in the second case people are prepared in the sense that they have already appealed to specific products even before they come into fashion. Shiller et al. (1984) proposition of social interaction and mutual interference can be traced back to the analysis of Katz (1957). Katz compared four studies to see whether the developed hypothesis of a two-step communication in society can be proven. The idea is that different groups, such as

families or friends, tend to follow one opinion leader, who is better informed by the mass media than others. This shows that social interaction is much more critical in the decision-making process than might be expected. Research has shown that opinion leaders may only lead in one field of expertise, but be influenced in another field by somebody else. Opinion leaders are not solely present in the better-educated segments of society; they are present in all segments.

Akins et al. (2011) found that information asymmetry has a healthy relationship to pricing and is further linked to the level of competition within the market. They assume that more substantial competition between informed investors leads to more transparent prices due to the high adjustment rate of prices to private information. In his article on psychological influence on investors, Hirshleifer (2001) also confirms the point of view that depending on the amount of available information it is hard to process all of it. The human brain is limited in its capacity. Habits are used as an argument for the repetition of individual behaviour because someone would have had a good reason to act in that way before. It seems that our brain is searching for more natural alternatives than processing and work. The same can be observed with the Halo Effect (Nisbett and Wilson (1977)), which shows that people tend to ignore rationales when one stock shows a currently good growth.

SUMMARY

Behavioural finance has provided the field of finance with many answers to observed anomalies and unrealistic assumptions. Behaviour and the way humans process information is influenced by routines, habits and social pressure. As Pressman (2006) argues, people follow behaviour because they have learned it and observed other people doing the same thing.

However, the field still lacks alternative theories which incorporate the human factor as a solution. Multiple areas, such as psychology and sociology, contribute to behavioural finance and provide new ideas regularly. The different attempts to define the field show that the research community is still not sure what precisely behavioural finance should be. Due to the lack of alternatives and the fact that none of the definitions has provided alternative models, it has to be assumed that behavioural finance will never be able to replace the neoclassical approach.

It should instead be accepted that behavioural finance has simply invited in other disciplines to show where the field has weaknesses and where improvements are needed. Behavioural

finance can, therefore, be seen as a way to introduce more reliable models. Hodgson (1998), for instance, described behavioural finance as evolutionary economics.

1.3.1 BEHAVIOURAL FINANCE IN REAL ESTATE

Due to the dominance of the classic finance theories, investors have applied these methods to real estate as an asset class to verify investment decisions. However, since the theories have not been initially developed for the real estate market, the application faces high barriers. Nevertheless, trades in the market are performed by humans, which are influenced by their perceptions. Therefore, behavioural finance has entered the real estate market. Kishore (2004) provides a comprehensive summary of behavioural research in the real estate discipline. In line with other authors, his summary leads him to the conclusion, that real estate markets are inefficient or at best only weak-efficient. Little is known about the influence of psychology and property investor irrationalities.

In Hardin's (1999) point of view, the real estate discipline adopted behavioural approaches relatively late. Other disciplines such as marketing or accounting used behavioural explanations earlier. One reason for the late acceptance can be the difference in the underlying object of interest, whereas marketing shows a stronger link between people's opinions and minds – real estate focuses on properties. However, Wofford et al. (2011) point out that early studies had already been done in the 1970s and mid-1980s by scholars like Ratcliff (1972) and Wofford (1985). Both these looked into the subject of behavioural finance with a focus on market participants and their cognitive abilities to process information in a decision-making process. Hardin (1999) examines the question of heuristics in the real estate market and how they narrow the available options down in a decision process. Other authors such as Northcraft and Neale (1987), Levy (1997) and Diaz (1997) also contributed to this question. The underlying idea is that people are likely to use anchoring when they have limited information about the subject. This can be observed, for instance, in the valuation process.

Gallimore (1996) clarifies that valuations are an essential field of behavioural research. A reason for this is the fact that valuations are proxies for prices and a function of information. More precise valuations are a function of how valuers process information. He conducted a series of interviews in order to identify whether values are subject to confirmation bias or not. Given several shortcomings, which are the result of qualitative research, the author concludes, that valuers are likely to confirm their opinions, instead of setting them objectively.

Other works of Wolverton (1996) and Gallimore and Wolverton (1997), did focus on the analysis of selection and confirmation biases of valuers. Here comparables, that match the assumed house price are more likely to be chosen, than properties, that might be an actual better fit. Another line of research has dealt with external influences by clients on the valuation process. Levy and Schuck (2002) found that price estimates are influenced by clients after the valuation took place.

Diaz and Hansz (2007) presented a comprehensive summary of behavioural research in connection with property valuation. According to the authors, valuers are subject to different forms of anchoring, for instance in the case, when they try to meet the expectations of their clients. Experiments have shown that valuers are also influenced by the information they are presented. Interestingly they tend to correct unrelated subsequent valuations upwards, in the case they know, that the previous valuation was below the contractual selling price.

MacCowan and Orr (2008) used a behavioural approach to explain why property fund managers dispose of specific properties from their portfolios. They showed that managers do act rationally, but are influenced by information which has been generated by irrational processes, such as biased valuations. The study shows that holding periods do shorten over time and properties are dropped because of portfolio restructuring. As another result of the study, it can be seen that managers base their decisions on external information such as market reports from real estate agencies. However, since markets are not fully transparent, managers are forced to base decisions on this biased information. Hardin (1999) made the further criticism that real estate should not only rely on the achievements in other fields but instead should develop field-specific explanations for individual behaviour.

Byrne et al. (2013) examined the U.K. property market and analysed whether it could be described as rational when using the underlying modern portfolio theory framework as a cornerstone of portfolio investment. They found that institutional investors show irrational behaviour in the composition of their portfolios. As a comparable measure for investable regions and property types within the U.K., the authors used an Investment Property Databank (IPD) dataset. This, however, might be influenced by the availability of assets, and institutional investors instead prefer to buy any property rather than none. Herding can be one reason for the significant variation within the portfolios in comparison to the suggestions of the dataset. Wofford et al. (2011) suggest that real estate portfolio managers should be aware of the limitations of human cognitive abilities and use this knowledge to improve the corporate structure and avoid such risks in the decision-making process. In another case study on the U.K.

market, French (2001) also focuses on the decision-making process of managers of pension funds during their asset allocation. The results imply that decisions are based on hard factual information like historical data, but they are also influenced by "current market perceptions and attitudes toward the real estate market". French takes this as proof that decisions are not entirely based on rational models.

Many scholars focus their analysis on the residential market. The advantage over the commercial real estate market is the frequency of trades and in some countries the data quality. Among others, Graham et al. (2007) analysed behavioural issues in the residential market. They explored whether catastrophic events such as hurricanes on the coast of North Carolina lead to irrational behaviour in the residential market. An increasing number of hurricanes in one region led to a shift in the willingness of buyers as to how much they wanted to pay and of sellers as to how much they were willing to accept. The authors observed an increase in the spread since buyers were afraid to face higher losses even though this fear is not justifiable.

Next to the analysis of the decision-making process in the real estate market scholars also focus on observed anomalies. One of the significant anomalies which can be observed in the market are calendar effects. This observation helps to disprove the Efficient Market Hypothesis since these regular patterns should not occur if market participants acted rationally. Different studies have shown that the real estate market displays this phenomenon. One of the first studies was undertaken by Brzezicka and Wiśniewski (2013). They showed that there is a July and an April effect, where the first one is influenced by fundamentals, but the latter can also be explained by a behavioural approach. For the intra-month effect, the authors suggest that market participants can control their market interactions according to this observation and improve their returns. However, those results should be treated with caution, since the analysis was performed only on one town in Poland. Also, the number of transactions was limited. Nevertheless, the authors conclude that behavioural influences are present in the real estate market.

Joel-Carbonell and Rottke (2009) extended the evaluation towards real estate investment trusts (REITs). This hybrid between real estate and stocks is influenced by behaviour to a more significant extent. The REIT market itself shows other advantages in comparison to the pure real estate market such as higher frequency and higher volume of trades. Joel-Carbonell and Rottke (2009) tried to prove whether the REIT market is affected by the IPO anomaly. They found that there is an under-pricing phenomenon in combination with an IPO. However, the authors believe that this does not naturally prove that the REIT market is irrational since not all investors

have the same chance of being allocated with shares at the beginning. Hui et al. (2014) followed the earlier warning not just to examine if there are behavioural anomalies, but also to examine whether those observations are consistent. They introduced two new tests to survey if the observed calendar effects have an economic impact. The suggested tests are White's Reality Check and Hansen's Superior Predictive Ability Test. The authors made the criticism that previous studies all rely on the same dataset and the same methodology. Hui et al. (2014) found that in many markets the December effect was statistically significant, whereas other effects such as the Sell-in-May effect were not. Furthermore, some effects seem to disappear over time. Given the new test, the authors were able to show that even the December effect had become economically insignificant. This would suggest that calendar effects do not play a considerable role and investors who are using such effects would not make better returns in the long run. What the analysis excludes is the possibility of a self-fulfilling prophecy, as was introduced by Merton (1948). The theorem states that a "false conception becomes true" when it leads to a change in behaviour. So, if many investors do believe that calendar effects are present in the market, they might become true. Another paper examines if there are any momentum effects in the residential housing market. Beracha and Skiba (2011) used metropolitan statistical areas in the USA and built zero-cost portfolios. They employed a long-short portfolio strategy and were able to generate abnormal returns. The authors surmise that the housing market is less efficient than other markets where more liquid institutional investors are present. This inefficiency is caused by transaction costs and the state of buyers and sellers.

Kaplanski and Levy (2012) applied psychological and medical results to the real estate market. They assumed, and this is in line with the results of Hirshleifer and Shumway (2003), that the mood of people is influenced by externalities. They analysed price changes in the USA, the U.K. and the Australian market, and linked them back to the change in hours of daylight and latitude. Even though this is no market-specific factor, it can be seen that externalities influence investors on all asset classes.

DeCoster and Strange (2012) looked at the behaviour of developers and how they reacted to the news on the market. The analysis shows that even when the market was supplied with the information of an approaching downturn developers kept on building. The main reason for this according to the authors is herding. Developers may be afraid that they will lose their reputation in comparison to other market actors when they change their behaviour and are proven wrong. On one hand, the efficient use of information could have protected the market as well as the developers; on the other hand, those market actors are not acting rationally at

all. What the authors exclude from their explanation is that developers may not have another chance to complete their buildings to minimize running costs.

SUMMARY

This short overview has reviewed where behavioural finance has reached the real estate market. Even though real estate counts as an alternative asset class, investors apply neoclassical theories to real estate investments, especially in a portfolio framework. As has been shown, the neoclassical approaches ignore the individual with his or her perceptions. The increasing literature on behavioural finance topics in the finance field and the real estate field indicate the interest and ambition of researchers to improve our understanding.

In general, the application and introduction of new methods to the real estate market are delayed in comparison to the equity market. It is surprising to see that early studies were performed during the 1970s. Nevertheless, scholars have to be careful with the transfer of behavioural finance ideas towards the real estate market. As criticized by Hardin (1999) the field needs to develop its own understanding of the relationships, due to market specifics, which differ from the equity market. Still, the broader research can be divided into the same two fields as suggested by Park and Sohn (2013). Scholars are likewise interested in the decision-making process of individuals and the formation of anomalies.

However, I have observed a tendency of research towards the housing and the REIT or real estate securities markets. There are no reasons given as to why researchers exclude the commercial real estate market from their analysis. Assumed reasons are the limited availability of data and the infrequency of trades. The studies of MacCowan and Orr (2008) and Joel-Carbonell and Rottke (2009) demonstrate that the real estate market is much more rational than may be assumed. Irrational behaviour influences information which is used in the decision-making process. This, on the other hand, leads to sub-optimal decisions and mistakes. Phenomena which are present in the equity market also occur in the real estate market but do not automatically lead to the acceptance of inefficiency. Some effects instead vanish over time or do not show economic insignificance, as proven by Hui et al. (2014).

Another aspect which is not discussed in the literature, but should be included, is the time frame difference in both the decision-making process and the investment period. In both cases, real estate focuses on a more extended horizon. This may give real estate investors more time to analyse information and to weight individual options more carefully. Following this

underlying assumption real estate should be less influenced by behaviour than the equity market since decisions are not made impulsively.

1.4 CHAPTER DESCRIPTION

The thesis is structured as follows. Chapter 2 will provide a comprehensive literature review of the state of the art in the field of sentiment analysis. The literature review will also shed light on the different methods which are used to extract the sentiment. Indirect methods and the use of sentiment proxies are of special interest for this thesis.

The third chapter utilizes the established methods for sentiment extraction and introduces a set of new sentiment indicators. The chapter investigates the commercial real estate market (office and retail) on a European scale. The sentiment indicators assume that even imperfect sentiment proxies carry some true sentiment. I also pick up recent developments in the field and use a composite indicator based on online search volume data to measure the underlying market sentiment. The different sentiment indicators are subsequently applied in a yield modelling framework. My findings suggest that more mature and probably more transparent real estate markets (i.e. Germany, France and the U.K.) rely to a larger extent on property specific sentiment, while less established markets have a stronger tendency to macroeconomic information. Reasons could be that property specific indicators do already incorporate wider macroeconomic information for those countries. On the other hand, do investors have to rely on all available information they can gather. Different to these assumed mature real estate markets, many East European countries don't have a large network of real estate service providers, which offer deeper market insight. The same accounts for functional REIT markets. While more mature real estate markets do offer these, many East European markets don't. This makes it difficult for foreign investors to get insight in the market. Therefore, investors need to rely on macroeconomic measures and draw their conclusions from here (please refer to chapter 3.6.5.4). The chapter will conclude with a summary of the key findings and a description of a range of shortcomings.

The next two chapters represent the crucial part of this thesis. I draw on the most recent developments in the field of sentiment analysis and apply natural language processing and textual analysis techniques to real estate documents. Due to the variety of markets and the novelty of the application, I have moved the focus from Europe to the U.K. and in particular to the London commercial real estate market.

Chapter 4 starts with an introduction to the field and provides a summary of the relevant literature, before illustrating the basic methodology for text pre-processing. Finally, I compare four different methods, which all share the same lexical methodology, where documents are categorized into either a positive, neutral or negative class, based on different word lists. The analysis is performed on a unique dataset compiled from market reports of all major real estate service agencies in the U.K. The results suggest that the use of sentiment indicators in a total return modelling framework provide useful information and improve upon the base model. Even in comparison with direct or the earlier constructed indirect indicators, the textual indicator produces significant results. The chapter concludes with a summary of these findings and an outlook as to where the applied method can be improved.

Chapter 5 illustrates a more advanced method which untightens some of the strict textual analysis assumptions and moves beyond the bag of words approach. Here, I use two new datasets. The application of various supervised learning approaches requires a training and a test dataset. Due to the absence of a labelled training dataset for the U.K. and especially for the commercial property market, I improvised by using *Amazon* Book reviews on real estate related books. In total, more than 200,000 book reviews have been used to train various algorithms. For the test dataset, on the other hand, I collected more than 100,000 news articles related to the commercial real estate market in the U.K. The developed supervised learning indicators were then used to extract the sentiment from the news articles. Results within a probit model framework are promising. The analysis with an unmodified method and unmodified text corpus only produced minor improvements in comparison to the lexicon approach. The supervised learning algorithms trained on book reviews fail to provide sufficient information. However, the combination of both methods provides a suitable bridge for the absence of a labelled training dataset, and the generated results are able to outperform other indicators with ease.

Chapter 6 concludes the thesis with an in-depth discussion of the findings. It is further pointed out where the research has limitations, and in which direction future research might head.

2 LITERATURE REVIEW

Sentiment analysis has been widely discussed in the academic literature. The field has its origin in the equity market and in consumer behaviour studies, where traders and other market participants tried to understand the underlying market sentiment.

2.1 SENTIMENT ANALYSIS

Chapter number one has provided a short introduction to the field of behavioural finance. Since the field has emerged in many different ways, it is necessary to place the focus on subcategories to get a better understanding. Sentiment analysis has always been used for behavioural analysis, and it has been adopted in a variety of other fields. Primarily through the intensive use of computers, sentiment analysis has become more and more popular. The extraction of sentiment is not only of interest to investors, who like to examine what noise traders do. Governments are also interested in this field since sentiment indicators provide insight into future economic developments and enable state institutions to prevent poor economic situations via the use of corrections.

In the next section of this chapter, it will be shown how sentiment analysis has emerged and what academics mean when they talk about sentiment. It is my goal to categorize the available sentiment indicators and to illustrate which methods are standard for extraction.

In the following section, I deal with the real estate field. What proxies have been used and what differences are present compared to the equity market?

2.1.1 SENTIMENT ANALYSIS ORIGINS

Sentiment describes an opinion, which somebody has or expresses. The word is derived from the Latin word *sentire* (feeling). Sentiment also describes a feeling or an emotion. Within the literature, a precise definition is not found. The term is used in different relationships. One definition states that sentiment analysis is related to textual analysis, where it is used as a synonym for opinion mining based on digital techniques to extract someone's attitude towards a specific topic or product. Bormann (2013) criticizes many of the following researchers for their

lack of willingness to provide an accurate definition of sentiment. His main point of critique is manifested in the argument that researchers try to explain the impact of sentiment on the market instead of explaining what they mean by sentiment. Bormann (2013) uses a psychological approach to define sentiment. In his opinion, short-term sentiment is equal to feelings and in the long run is more equivalent to the mood of market participants. This, however, can be seen as wordplay, since the author only changes the underlying meaning, but does not offer an in-depth definition himself.

In the economics literature, sentiment analysis plays a huge role. Scholars are motivated by the observation of herding behaviour. With a deeper understanding of the underlying sentiment of investors, models and predictions about the market movement could be improved. A broad definition of sentiment from a financial point of view can be found in Baker and Wurgler (2007) where sentiment is the belief of investors about future cash flows and investment risk that is not justified by the facts at hand. The authors further state that betting against sentiment is costly and risky, which is why arbitragers hold off on their actions.

The academic literature can be sorted into two main categories of sentiment measures: market-based measures and survey-based measures. According to Hengelbrock et al. (2013), the market-based measures include, among others, closed-end fund discounts, liquidity figures and trading volumes of the underlying asset. Other proxies are based on interest rates, labour income or GDP figures. It is assumed that those proxies provide enough insight in the market or the underlying asset and its behaviour. Transaction-based measures, for instance, allow a conclusion on the popularity of an asset, given the trading volume. Other factors, such as macroeconomic variables, are unable to shed light on an entire market, individually. Survey-based measures extract the sentiment either in a direct way with the help of interviews or in an indirect way where the opinions of market participants is expressed in newsletters. In general they do not require any further modification, in order to extract the sentiment.

Following this motivation, many scholars try to find a suitable proxy for the sentiment of investors. Among others, Barberis et al. (1998) applied psychological ideas to their model. They focused on the phenomenon of over and under-reaction and simplified the environment of their assumed traders, who will be risk averse and only operate in two different regimes dictated by their economic environment. They based their model on the observation that news is only slowly incorporated into prices. However, the authors left the reader without a real-life application of their model.

Lee et al. (1991) have shown that sentiment does play a role in the financial market. They have analysed closed-end funds and their exposure to noise traders. Those funds have been traded with discounts which can be assumed to be an indicator of the expectations of the traders for future development of the asset. The more significant the exposure of the fund, the more sensitive is the discount to the investor sentiment. Even though the authors performed a wide-ranging analysis of this relationship based on the correlation of the discounts and the returns of the underlying stocks, they have been at the centre of some criticism. Elton et al. (1998) examined that the suggested closed-end fund sentiment index by Lee et al. does not enter the return generating process more frequently than other indices. They further run a counter-experiment with a focus on companies where the majority of shareholders are institutional investors. The assumption is that those companies are less sensitive to investor sentiment. They were able to prove that the industry measures are competitive with the sentiment index.

Baker and Wurgler (2006) reached a similar conclusion as Lee et al. (1991). Although they did not focus on closed-end funds, they found that investor sentiment has a more substantial impact on the returns of small, young and highly volatile stocks. The researchers were able to show that returns are higher (lower) when sentiment is weak (strong) at the beginning. This is logical since stocks which experience high sentiment have already higher attention and usually higher prices, which would reduce the margin of returns. In the same year, Kumar and Lee (2006) used an extensive dataset of retail investor transactions to prove that investors buy and sell stocks in concert. Since this trading group is more likely to focus on small, young and highly volatile stocks, the findings are consistent with Baker and Wurgler (2006), and later further confirmed by Liang (2016), Aissia (2016) and Frugier (2016).

Scholars such as Brown and Cliff (2005) contributed to the broad field of sentiment analysis, using the sentiment index of Investors Intelligence[1]. This proxy is based on the textual analysis of a number of market newsletters. The authors included further control variables in their model to examine the actual impact of the sentiment proxy; among others, they used the US Treasury Bill and US inflation rate. Due to the incorporation of the sentiment index, the authors were able to predict market returns over a three-year horizon and showed that irrational behaviour does have an impact on asset price levels.

[1] Investor Intelligence is a UK based data provider. Data Is provided on a subscription base. The service is offered for more than 50 years.

Sentiment analysis is not only performed in the equity market. Even earlier, researchers such as Katona (1968) tried to understand consumer behaviour. They analysed sentiment within the society of consumers via the use of surveys. As one of the leading sentiment indices, the University of Michigan Consumer Sentiment Index emerged in 1947 based on the remarkable work of Katona. Ever since the index was established, researchers have used the index for predictions for the US economy. Among others, Carroll et al. (1994), who tried to explain how the index predicts US household spending, found a positive correlation between lagged values of the index and lagged values of consumption. However, the evidence suggests that the index can only explain current relationships rather than future developments.

Based on this work, Bram and Ludvigson (1997) and later Howrey (2001) compared the index to the Conference Board Consumer Confidence Index. Bram and Ludvigson argued that the partial focus on the Michigan Index in many academic papers may not fulfil its purpose, in the sense that it is not clear whether the predictions about future spending of consumers actually hold. In addition to this, the authors questioned whether the prediction of confidence indices might not already have been incorporated in other economic benchmarks. Both indices are based on five questions, whereas the Conference Board Index has two specific questions which are aimed at the opinion on the current job situation. The authors demonstrate that those questions do have a higher educational value about future consumption. In the case where multiple sentiment proxies are used at the same time, it should be considered, that many aspects of the two consumer indices are already covered by other benchmarks such as interest rates or labour income. While the consumer indices only provide a marginal insight into what the drivers of the consumption are, those hard facts, actually provide a direct linkage.

Howrey (2001) showed that the Michigan Index alone, as well as in conjunction with the Conference Board Index, was able to predict GDP growth one quarter ahead. Other scholars such as Dominitz and Manski (2004) have pointed out that consumers lack experience about economic relationships and that their opinions should be treated with caution when it comes to predictions. Frugier (2016) has pointed out that in general a range of different sentiment proxies is used. However, they seem to be highly correlated.

Due to the fact of strong linkage of the above-mentioned indices (Conference Board Index and the University of Michigan Consumer Sentiment Index) to US economy, Easaw and Heravi (2004) run their analysis for the U.K. market with the help of the Consumer Confidence Indicator provided by Gesellschaft für Konsumforschung (GfK). Their results were similar to Bram and Ludvigson (1997). The predictive capability of this index for important consumption goods was significant. However, it seems that cultural or economic reasons also influence the power of the

predictions of those indices. Either due to this or due to the different structure of the questions of the survey, Fan and Wong (1998) were unable to prove the findings of Carroll at al. (1994) for the Hong Kong market. In addition, Malgarini and Margani (2007) looked at the Italian sentiment index and showed that the Italian market is predictable. They identified that different consumer groups are differently affected by economic and political shocks, such as elections. Another study by Hung (2016) used consumer confidence as a sentiment proxy for the Taiwan stock exchange. In the author's opinion, the forward-looking element of the index is used to capture future behaviour.

Another problem which arises from regional differences is the increasing trend of cross-sectional and multi-asset investments. Froot et al. (2014) tried to find a suitable solution to cover general sentiment in multiple markets and for multiple asset classes, including U.S. equities, U.S. real estate, bonds and commodities. The broad sentiment indicator that the authors developed is called a behavioural risk scorecard which covers different specifics (i.e. sign, momentum and direction). They showed that the use of the scorecard could improve investment decisions since the risk can be better estimated and investors have a broader insight.

All these examples show that it is possible to examine the sentiment of people. Yet, there are country specifics, meaning that each country may has their own current economic development, which differs even in larger economical circles such as the European Union. In addition, the predictions for the current situation are much more accurate than the predictions of the future. And likewise, existing benchmarks such as interest rates or labour income may cover the influence of consumer confidence indices in a better way. One reason could be, that the national trend is incorporated in those indices and that consumer confidence is just a mere aggregation of these factors. Following the achievements of behavioural researchers, the incorporation of the human factor in models helps to improve our understanding. But still, the majority of these examples is based on sentiment indices which are computed from surveys. So, there is a high barrier to obtaining access to the sentiment of traders or consumers. Sure, it might be possible to use existing sentiment indices, yet not all countries have a sentiment index, and the computation is long lasting and probably financially intense.

Therefore, researchers and market participants have sought to find other ways to extract sentiment. Search engines such as Google provide free access to the search queries of millions of people. Search engine data has been identified by many scholars as a source of sentiment. Since Google search entries represent the attention and interest of individuals, who are the smallest unit of the economy, it is possible to draw general conclusions from here. However,

many searches on Google are only interest-driven and do not automatically translate into a specific action in the stock or property market. A different point of view is presented by Barber and Odean (2007), who follow the belief that investors only buy those stocks, which have caught their attention. Meaning, that there must be an initial factor or influence, that provoked the interest. Following that, one could argue that Google serves as a medium to increase the knowledge of an investor to whom a new investment has been brought to attention. However, even in this scenario the aggregation of all searches would allow to get some idea about the market interests. The reader should keep in mind, that Google search entities are only used as a sentiment proxy since it remains unknown what the intentions of the searcher are.

Among others, Joseph et al. (2011) used stock ticker symbol searches on Google. The developed sentiment proxy based on the intensity of the searches was able to predict abnormal stock returns as well as volume. According to the authors, those results are consistent with the earlier achievements of Baker and Wurgler (2007).

One of the significant applications of Google Trends can be found in Ginsberg et al. (2009). The authors were able to show that nowadays behaviour has changed so much that it becomes traceable. People having the flu do start to look for symptoms before they go to a doctor. This finding is significant since it enables governments and health institutions to prepare for an outbreak. The authors were able to use Google Trends to see where the outbreak begins and how the flu spreads over the USA.

Using a social application of Google trends, Preis did some ground-breaking work. He was one of the first scholars who saw the potential and linked the tool to behavioural finance. In 2010 Preis, Reith and Stanley analysed the complex dynamics of the economic life, by linking Google search queries to the U.S. stock market. From the authors' point of view, the individual represents the smallest unit of the economy and provides millions of search queries every year. Those search queries reveal what people think and want. The authors linked weekly transaction volumes of companies in the S&P 500 with the corresponding search term on Google. Both time series are correlated. It was observed that an increase in transaction volume goes along with an increase in search volume and vice versa. The authors were unable to see any preference in an increase in searches and whether the company was bought or sold. This is why they assume that news and volume are strongly linked together, since its presence in the news can be a trigger for an increased search.

In 2012 Preis, Moat, Stanley and Bishop extended the previous research on Google Trends. They made clear that the amount of available data and information had increased over the

previous couple of years. The authors point out that significant data sources provide enormous possibilities for behavioural studies. Their paper analyses the cross-country behaviour of inhabitants as to whether they are future orientated instead of focused on the past. The reason for such an analysis is to prove that the internet and the handling of major economic events have changed over the years. Countries with a higher GDP per capita do have inhabitants who will be much more interested in the future based on Google Trends data.

In 2013 Preis, Moat and Stanley looked at Google Trends data on the trading behaviour of individuals. The underlying assumption is that the interaction between individuals and the internet can give early warning signs of significant stock market movements since the searches on Google do not only reflect the current situation on the stock market but provide signs of future developments. This assumption is based on the research work of Herbert Simon, who assumes that actors begin their decision-making process by gathering information. In times when market participants have stronger concerns before they invest, the authors assume that searches on Google increase. Preis et al. developed a trading method based on 98 search terms which are partly suggested by Google's related words. Based on the weekly change in stock end prices of the S&P 500 and the changes in the correlated search terms provided by Google Trends, the authors sold a composite of the Dow Jones index when the search volume increased for specific terms, such as "debt", and the other way around. Following this method, the authors were able to generate a significant profit in comparison to a typical buy and hold strategy.

Similar research was performed by Choi and Varian (2012) when they analysed a series of different economic fields such as house sales. Contrary to Preis et al. (2012) they do not support the assumption that Google data can help to predict the future but not the present. This result is consistent with Fuhrer and Wilcox (1994) who also confirmed that predictions of the present are more accurate than predictions of the future. Vosen et al. (2011) picked up the initial work of Choi and Varian (2012) and focused more on the consumption of U.S. households. They compared a constructed Google Trends indicator with the University of Michigan Consumer Sentiment Index and the Conference Board's Consumer Confidence Index. Their results suggest that the online search volume-based index is able to outperform the other two indices in terms of forecast accuracy. The researchers applied a simple autoregressive framework. They conclude that Google data is able to forecast consumption within the USA.

Loughlin et al. (2014) combined Google Trends with the Twitter-like application StockTwits to analyse herding behaviour. They pointed out that ground-breaking work from Bollen et al. in 2010 had proven that social media applications can help to increase the prediction of the stock

market. The authors used the more finance orientated Twitter-like application. With a focus on stock returns of just four stocks, Loughlin et al. (2014) did not find a significant correlation between Google Trends and the stock returns, whereas the generated index from StockTwits showed a sure success.

A similar approach was taken by Sprenger and Welpe (2014). They analysed StockTwits as a significant source of information for experts and individual traders. Their results show that microblogs such as Twitter can be seen as a reliable and comprehensive source of information for financial trading.

SUMMARY

It is without question the case that sentiment is an essential factor in market influence. However, the critique of Bormann (2013) is legitimate. In most of the presented academic papers, a definition of sentiment is absent. It seems that researchers have somehow agreed on a definition, which could rely on psychological terminology, since the field is strongly related to behavioural finance.

Among others, Baker and Wurgler (2007) showed that academia had been ignoring the issue of whether sentiment influences the returns of stocks or not. Academia is now investigating how sentiment should be measured and interpreted. This angle was picked up by a variety of researchers, who showed that sentiment based on surveys or even based on Google search volumes may only help to predict the present rather than the future.

So far sentiment is either based on a range of macroeconomic proxies, or it is based on surveys, which are not present in all countries. This limits the work of researchers as well as the work of market participants.

Even when markets do have a sentiment index, results cannot be transferred from one market to another, as shown in the example of Hong Kong or Italy. It seems that culture has an impact on the predictions of sentiment indices.

As mentioned, the work based on online search engines is auspicious. This new approach, which is based in-between surveys and sentiment proxies, reveals the thoughts of millions of people. This is interesting from both points of view: that of consumer behaviour and of retail trader analysis. Access to specific searches can be seen as the combination of surveys and proxies.

2.1.2 SENTIMENT ANALYSIS IN THE REAL ESTATE MARKET

Real estate is a significant asset class and as one of the most significant consumer goods of the society and it has not been excluded from the analysis of sentiment. The financial crisis in 2007/08 sets the focus of sentiment analysis on the real estate market. The motivation of market participants to discover the underlying drivers of noise traders are similar to the intentions of equity market participants. Essentially, there are three factors which should be considered in order to understand the sentiment within the real estate market. First, in which market is the transaction situated? Second, who are the market participants? And finally, how much information is available during the transaction process?

Researchers divide the market into a private and a public real estate market. Both sides do have their own requirements and ask for different sentiment measures. Public markets are much more liquid and transparent. It is unclear, if there are noise traders in the private real estate market, who can benefit from these market requirements. Real estate is a long-lasting and intense capital investment, speculative and irrational investments are much more seldom compared to equity investments. The frequency of trades and the rationale behind them can be assumed to be different, at least in some parts of the real estate market. Irrational behaviour in both the private residential and commercial real estate market can be triggered by specific developments in the market. Private investors may be afraid that they will not be able to enter the market at a later stage when prices increase. The same applies to institutional investors, who may be attracted by new developments or trends which could lead to irrational decisions. So, a specific group of noise traders might not exist, but irrational thinking motivated by external factors can be assumed.

Another factor, which does influence the scale of sentiment, is the availability of information. Publicly traded assets are assumed to have a greater information coverage and investors are less uncertain, when it comes to predictions about the market. Yet, private markets suffer from information asymmetry. It is more costly to gather all information, which are needed to make a sound decision. At the end of the process, this leads to better-informed investors in the private market. The absorption of shocks in the sentiment, however, takes longer, due to the lower frequency of transactions and the accompanying fact that prices are not documented continuously. Private real estate markets are therefore, stronger influenced by market sentiment.

Further, differences arise when different asset classes are examined. It is assumed, that the residential market, for instance, absorbs sentiment shifts much faster than the commercial real

estate market (i.e. Nanda and Marcato, 2016). Reasons are, that the number of transactions is much higher in comparison to the CRE market. That allows a more rapid conclusion about the market development.

Case and Shiller (1989) tried to find proof that the housing market is inefficient or at least less efficient, compared to the financial market. They were motivated by the observation that prices and returns are more like a random walk than logical patterns. Another reason is that the market is dominated by individuals, who privately trade their houses they live in. This observation was underlined by the fact that changes in interest rates are not absorbed by real estate prices. Colossal data issues do prevent final and general results. The authors were unable to prove markets either to be inefficient or efficient, due to the individual characteristics of the market. Their results show that the market is non-transparent and possibly driven by irrationalities.

STUDIES ON PRIVATE REAL ESTATE MARKETS

Similar to the above-described sentiment analysis, the general separation of the applied measures in the literature remains. Scholars use survey-based sentiment analysis and market-based analysis with the help of market proxies for the examination of market sentiment.

Goodman (1994) made the criticism that many of the published survey-based indices are privately funded. He does not explicitly point out that institutions may enter a conflict of interest, but his criticism at least should lead to a higher awareness. He further analysed three survey-based indices for their short-run forecasting power of housing statistics, such as housing starts, and new and existing home sales. The intention behind his analysis is based on the fact that those surveys are published weeks before the hard statistics. Goodman concludes that the forecast results are minimal in the short run. However, his analysis lacks full depth, and the author somehow excludes long-run trends or even the possibility of lagged values.

Case, Shiller and Thompson (2012) looked at the financial crisis with the help of survey data, which has been collected over a 25-year horizon. They criticize the lack of research regarding the expectations of home buyers before and during the first stages of the crisis. They assume that insight into the thought processes of home buyers may help to reveal why they bought a house during a crisis. The data reveal that buyers were aware of current developments, and in most of the cases, they acted correctly in the short run. However, their expectations, in the long run, were tremendously wrong. A similar critique towards the lack of research regarding the

thought process and the expectations can be found in Foote, Gerardi and Willen (2012). They provide a comprehensive analysis of the ongoing discussions, theories and reasons as to why the market was somehow healthy in its fundamentals, but that everybody was delusional and expected the market to develop as it has over recent years. The authors conclude that it is impossible to prevent bubbles when expectations in the whole market are positive.

Tsolacos (2012), analysed the application of sentiment indicators on the European private commercial real estate market. He pointed out that sentiment based on a survey level can be seen as the beliefs of market participants of future development, which makes sentiment an attractive feature in a forecasting framework. He used the economic sentiment indicator (ESI) provided by the European Union for three major markets in Germany, France and the U.K. The ESI is a combined indicator of four business surveys and one consumer survey. Adopting a probit model to the question whether it is possible to forecast turning points in three main office centres in Europe, the author revealed that the model is capable of giving early warning signs.

Dua (2008) cannot be sorted into one of the two above categories. She used proxies as well as survey data to prove her assumption that house buying attitudes in the USA are, among others, correlated with interest rates, wealth and housing prices.

Croce and Haurin (2009) were interested in the turning points of privately held residential real estate markets in the US. They acknowledged the importance of the estimation of these points for market participants on all sides: buyers, sellers and policymakers. They used the Wells Fargo/ National Association of Home Builders (NAHB) Housing Market Index and the University of Michigan Survey of Consumers index as to whether a time is right to buy or not. They were able to verify a statistically significant correlation between the two indices. To capture the market, they used housing starts, home permits and new house sales. In a comparison test, the Michigan Index outperformed the Housing Market Index (HMI) and is therefore favoured by the authors for predicting turning points. However, the authors further note, neither of them has produced entirely satisfying results.

Jin, Soydemir and Tidwell (2014) extended the work of Croce and Haurin (2009). They identified that a sentiment factor might be suitable to predict price changes in the US housing market. Instead of using the HMI, they decided to use the Case and Shiller House Price Index and the Conference Board Consumer Sentiment Index. With the help of error correction models, they were able to show that house prices are correlated with the underlying sentiment of the market. Similar to Baker and Wurgler (2006/07), the authors decided to orthogonalize imperfect fundamental market proxies.

Clayton, Ling and Naranjo (2009) picked up the fundamental issues of real estate markets. They referred to non-transparency, illiquidity and robust segmentation of the market, which all goes hand-in-hand with information inefficiency. Furthermore, investors are unable to short sell the asset, which all leads to a sentiment-influenced market, with a strong bias to mispricing. Their analysis of the commercial real estate market showed that the sentiment of investors influences the market even after controlling for changes in rental growth.

In a later study Ling, Naranjo and Scheick (2014) kept focusing on the short sale constraints in private real estate markets. The resulting hypothesis was that sentiment has a much stronger influence on private real estate markets than the publicly traded real estate markets, due to the fact that market or price correcting mechanisms do not work. The authors used both direct and indirect measures of market sentiment, and they applied the methodology introduced by Baker and Wurgler (2006/07). They used eight indirect measures of market sentiment, following the idea that all imperfect proxies at least contain an individual share of pure sentiment. Ling et al. (2014) showed that prices and returns are affected much longer by sentiment shocks in the private market.

Beracha and Wintoki (2013) extend the work of Preis et al. (2013) and Choi et al. (2012). They identified Google as an optimal source of consumer sentiment and used the search volume as a proxy. The authors analysed whether the search volume on a US city level is able to predict abnormal price developments in the private residential real estate market. Since the real estate market is unable to adjust to changes on the demand side in the short-run, the correlation between search volume and price developments is high. The difficulty lies in the choice of search terms; it needs to be broad enough to be related to the intention to buy a property. The authors were able to show that search engine data can be used as a sentiment proxy for the housing market and price developments.

A large body of literature focuses on the USA and the private housing market (Choi and Varian, 2012; Da et al., 2011 and Beracha et al., 2013). Hohenstatt and Kaesbauer (2014) have focused on the U.K. housing market and have, among other things, shown that sub-categories supplied in the Google Trends tool are more suitable than a broader search volume index (SVI). The authors used the “real estate agency” sub-category to extract consumer sentiment in order to predict the transaction volume of privately held houses. Further, in Das et al. (2015b), the authors have been able to link search queries to market fundamentals and showed that an increase in searches for rental apartments corresponds to a decrease in vacancy rates.

Similar research was completed by Dietzel, Braun and Schäfers (2014). They constructed three different proxies based on the Google search volume. Once more focusing on the U.S. market, the authors showed that it is possible to apply sentiment analysis to the private commercial real estate market. They used the CoStar Commercial Real Estate Repeat-Sales Index for a Granger causality test. Results reveal that Google search volume data is able to predict the market. However, and this is consistent with other studies, the authors suggest that better results are achieved when researchers try to nowcast rather than forecast. The authors criticize the same issues as do other researchers. Even though the tool is easy to use and free of charge, the lag of absolute search values and the data scaling leave the user wondering.

Baker and Saltes (2005) contributed to the literature via focusing on the commercial market. They used architecture billings in the USA as a leading indicator of construction activity. They point out that not all architectural activity transforms into construction activity. The constructed index was able to represent half of the market development and was capable of showing turning points. Conforming the criticism of Goodman (1994), the authors have to be marked as representatives of the private market. Furthermore, the authors point out that the data quality is poor. The used time series is shorter than one decade, and the data is not published on a frequent base.

Marcato and Nanda (2016) have analysed a range of sentiment measures. Confirming other results, they were able to show that sentiment measures help to forecast changes in private commercial and residential real estate returns. With a 20-year horizon of US real estate data, the authors applied a vector autoregression framework. However, the results are more promising for the residential market than for the commercial market. The authors assume that the latter one is not reacting as strongly as the residential market to shocks in exact sentiment. The authors also applied the above-mentioned method of Baker and Wurgler (2006/07). Among others, Marcato and Nanda used the University of Michigan Index, as well as Architectural Billings Index (ABI) (introduced by Baker and Saltes (2005)), and the HMI.

STUDIES ON PUBLIC REAL ESTATE SECURITY MARKETS

Sentiment analysis has been further applied to public real estate securities (REITs). Some of these studies, such as Barkham and Ward (1999) and Chiang and Lee (2009), use the traditional understanding of closed-end fund discounts as a sentiment proxy. Lin et al. (2009), on the other hand, draw a subtle distinction and illustrate that REITs behave differently to closed-end funds;

therefore, a separate examination is needed. They develop a sentiment measure based on the ownership share of REITs.

Barkham and Ward (1999) contributed to the question of noise traders in the public real estate securities market. They picked up the analysis of closed-end funds from Lee et al. (1991) and looked at real estate companies in the U.K. They showed that closed-end real estate funds are traded with a discount on average as well. This is caused by the noise traders who overestimate value changes in the underlying asset. The authors identified two groups of noise traders: stock investors and developers who are responsible for overbuilding.

Among others, Das et al. (2015) investigate whether a sentiment component can improve a REIT trading strategy. Rather than using indirect sentiment proxies, such as the closed-end fund discount, the authors use a survey-based measure for institutional investor sentiment. This is in line with the recommendation in the literature (Ling et al., 2014 and Lin et al., 2009) and their results suggest that a direct measure is superior in comparison.

In Freybote and Seagraves (2017), the authors first pick up on the idea of disaggregated sentiments for different investor types. Unlike previous studies, they define their sentiment measure as the general attitude towards the office market, expressed in trading behaviour. Following the idea of Kumar and Lee (2006) that noise traders trade in concert, the authors show that multi-asset property investors use the sentiment change of specialized property investors to adjust their trading strategy.

Freybote (2016) further underlines the predictive power of forward-looking sentiment measures. Using credit ratings or real estate specific indices results in the fact that backwards-looking elements dominate. A prediction of market movements is therefore limited.

Another sentiment proxy is the investor risk appetite in the public real estate securities market. This measure was introduced by Hui, Zheng and Wang (2013). They assumed that risk appetite would increase when market fundamentals are stable and positive and vice versa. The authors assume that investors do have their own specific risk appetite and do not change it regularly.

SUMMARY

This review has revealed that the real estate market provides enough evidence for sentiment driven developments. Researchers have not left any field untouched when examining whether sentiment influences the markets. Nevertheless, this overview also shows that real estate is much more bounded by its market characteristics. Lumpy investments, illiquidity and short sell constraints are only a few examples, which force researchers to be innovative to find suitable ways to examine sentiment.

With regards to the specific sentiment measures the literature has provided a series of different options. Publicly traded markets allow conclusions about the sentiment by utilizing information about REITs. In Ling et al. (2014), eight different indirect sentiment proxies were used (i.e. REIT stock price premium to the Net Asset Value (NAV), the percentage of properties sold each quarter from the NCREIF index, the REIT share turnover, etc.). Private markets on the other hand require more farfetched sentiment proxies, since the markets are not entirely dominated by professionals, here consumer spending and other macroeconomic factors play a crucial role. Private individuals have a different mindset by trading their homes they live in (Case and Shiller, 1989). It becomes clear, that a generalization of sentiment measures about entire markets and asset classes is nearly impossible. Surveys for instance are directed towards a specific market, either stated in the questions or through the participants. The point of view of how the market sentiment should develop depends on the investor class, which should be examined. For instance has a private investor a different sentiment when prices rise than a property vendor or developer. It remains questionable, if the sentiment of two opposing investor groups is the inverse function.

The general separation into survey-based measures and proxy-based measures remain in the real estate literature, but the impression occurs that researchers use both measures in a connected way, when it is possible. Orthogonalization, as introduced by Baker and Wurgler (2006, 2007) has been identified in both fields as a suitable method to extract sentiment from a series of imperfect proxies.

Giacomini (2011) gives a list of suitable sentiment indicators. For the general economy, the author mentions the University of Michigan Consumer Sentiment Index, the Conference Board Consumer Confidence Index and the Economic Sentiment Indicator provided by the European Commission. However, this list is far from comprehensive. For the classic stock market, sentiment proxies such as liquidity, mutual fund flows, retail investor trading activities and closed-end fund discounts, are listed. The authors mention in the private real estate market

commercial mortgage flows, the percentage of properties sold from the National Council of Real Estate Investment Fiduciaries (NCREIF) index, transaction activities and total return figures from transaction and appraisal-based indices, as suitable proxies. For the public real estate market, the author extends this list with the number of REIT IPO´s, average REIT stock price premium divided by the NAV and the net commercial mortgage flows.

As well as these specifics, the review has revealed that the majority of researchers keep on focusing on the USA and on the housing market. Among a few, Marcato and Nanda (2016) tried to apply their analysis on both real estate markets, but concluded that shocks in sentiment lead to stronger reactions in the housing market, which result is in line with other findings. Tsolacos (2012) focused on the European market and was able to prove that sentiment influences the office market. The housing market is characterized by a higher frequency and higher volume of trades. Therefore, the market is assumed to be able to adjust in a better way; however, it also shows stronger reactions to sentiment shocks. A reason for the focus on the USA might be the large amount of available research on sentiment indices. Nevertheless, this shows that the commercial real estate market in Europe is still under-researched. Based on the results of Tsolacos (2012), I think that sentiment factors also influence commercial real estate markets and participants. Therefore, the following analysis of this thesis proceeds with a focus on commercial real estate.

3 SENTIMENT PROXIES[2]

3.1 INTRODUCTION

The literature review has shown that the real estate market is influenced by sentiment in various ways. Researchers have focused on both direct and indirect sentiment proxies to measure underlying market sentiment. In this chapter, I have followed the general assumption that the underlying sentiment can be mirrored with the use of sentiment proxies.

However, different to other studies I will not look at either the USA or the housing market. Even though the results in this chapter support earlier findings, it is my intention in this first section to display the shortcomings of the standard approaches.

This study has a broad geographical coverage. The sample consists of important commercial real estate markets in 24 European countries and 48 cities. Cities such as London or Paris have been recorded with multiple regions (e.g. London City, London West End) in the dataset. Therefore, the total number of recorded regions is a total of 80 city regions (see Table 3:1). The data has been provided by Cushman & Wakefield.

[2] The main parts of this chapter have been made into a journal paper, which is currently under revision by the Journal of Real Estate Research. The title of the paper is "Which Sentiment Indicators Matter? An Analysis of the European Commercial Real Estate Market" by S. Heinig, A. Nanda and S. Tsolacos.

Table 3:1 - List of all countries and city-regions

Countries	City-regions	City-regions	City-regions	City-regions
Belgium	Amsterdam	Istanbul - Asian CBD	Newcastle	Stockholm
Czech Republic	Antwerp	Istanbul - European CBD	Nottingham	Tallinn
Denmark	Arhus	Kaunas	Oslo	The Hague
Estonia	Barcelona	Klaipeda	Paris (20 districts)	Triangle Area
Finland	Berlin	Krakow	Paris (CBD)	Utrecht
France	Birmingham	Kyiv	Paris Center West included CBD (1-2-8-9-16-17 districts)	Vilnius
Germany	Bristol	Leeds	Paris (IDF)	Warsaw
Hungary	Brussels	Liege	Inner Eastern Suburbs (Paris)	Zurich
Ireland	Bucharest	Limerick	Inner Northern Suburbs (Paris)	
Italy	Budapest	London	Inner suburbs (total northern, eastern & southern suburbs) (Paris)	
Latvia	Cardiff	London (City)	Inner Southern Suburbs (Paris)	
Lithuania	Copenhagen	London (Docklands)	Paris Left Bank/Bercy/ Gare de Lyon (12 & 13 districts)	
Luxembourg	Cork	London (Heathrow)	Paris (La Défense)	
Netherlands	Dublin	London (Midtown)	Outer suburbs	
Norway	Dusseldorf	London (WE)	Paris - Western Crescent	
Poland	Edinburgh	Luxembourg	Paris - Western Crescent - Neuilly Levallois	
Romania	Frankfurt	Lyon	Paris - Western Crescent - Northern Boucle of Seine	
Russia	Galway	Madrid	Paris - Western Crescent - Southern Boucle of Seine	
Spain	Geneva	Malmo	Paris - Western Crescent - Suburbs of La Défense	
Sweden	Glasgow	Manchester	Prague	
Switzerland	Gothenburg	Marseille	Riga	
Turkey	Hamburg	Milan	Rome	
U.K.	Helsinki	Moscow	Rotterdam	
Ukraine	Istanbul	Munich	Sheffield	

I have developed a set of four different sentiment indicators using principal component analysis and orthogonalization procedures. In addition, I present a more diversified sentiment indicator based on online search words at a regional level. The sentiment measures are tested in a standard yield model and a panel data framework. The quarterly data ranges from 2004q1 to 2014q4.

This study contributes to the literature in three ways. First, I confirm that sentiment can be extracted from indirect sentiment proxies. Four indicators are constructed that represent the irrational or unexplained aspect of market participants. These implicit sentiment indicators show a moderate correlation with direct sentiment indicators. Second, my findings show that yield models benefit from the explicit inclusion of sentiment measures. For both office and retail, the majority of models incorporating sentiment outperform a standard (benchmark) yield model on the basis of goodness of fit and forecast evaluation tests. Finally, the results suggest that real estate markets are more reflective of sentiment in less stable environments, a finding in line with the expectations. The reaction of investors in countries or markets with a limited amount of information and low liquidity can be vivid and impulsive since views formed about market developments are based on limited datasets. This finding is similar to the results from the closed-end-fund market or the stock market literature, where more permanent funds or companies react less to shifts in sentiment (i.e. Lee et al. (1991) and Lin et al. (2009)).

The next section of this chapter briefly summarizes the standard literature on yield models. The constructed sentiment indicators enter a standard yield model with the objective of improving the predictability of the dependent variable. Property yield is assumed to react to changes in the market more rapidly than rents.

The sentiment indicators, are based on both direct and indirect sentiment proxies. In general, I have followed the suggested method of Baker and Wurgler (2006, 2007) and used a principal component analysis and an orthogonalization process for the extraction of the sentiment. Besides the more established methods, another indicator based on online search volume data is used to measure the sentiment. To anticipate any critics at this point, who might question the choice of sentiment proxies, I have adopted the opinion of Baker and Wurgler (2006, 2007) that any imperfect sentiment proxy, at least to a particular share, carries some true sentiment.

The remainder of the chapter is structured as follows. The theoretical underpinnings will be discussed, followed by a description of the data and the methodology, before the results and

several robustness checks are presented. The chapter concludes with a summary of my key findings.

3.2 LITERATURE REVIEW ON YIELD MODELLING

There is plentiful academic research on the topic of the determination of cap rates or yields. Yield is the ratio of net operating income generated by a property asset over its price. Expected growth in net income from the real estate asset is one of the fundamental determinants of yields. Two widely used methods to measure expected income have been put forward by Hendershott and MacGregor (2005a) and Chervachidze and Wheaton (2013). According to these methods market participants form expectations on the basis of rent deviations from a sustainable or equilibrium path of rent. These deviations are seen as a suitable proxy for the expectations of market participants about near future rent movements that will impact on cap rates. Hendershott and MacGregor (2005a) view the deviations as a mean (or equilibrium) reverting process to which real estate yields respond. This argument finds empirical support in the U.K. property market but not in the USA (Hendershott and MacGregor, 2005b).

Sivitanidou and Sivitanides (1999) argue that the rent variable is likely to be the only component that carries locally fixed and time-invariant elements. Sivitanides et al. (2001) use panel data analysis drawn from the National Council of Real Estate Investment Fiduciaries (NCREIF) dataset and introduce two measures for the expected income growth: expected economy-wide inflation and expected real-rent growth.

Empirical investigations of cap rate movements attempt to incorporate the impact of the changing risk premium, its components and other national or local influences (economic and investment market) on yields (see Chervachidze et al. (2009), Chervachidze and Wheaton (2013), and Duca and Ling (2015)). Risk premia encompass a range of influences on yields including investor confidence and sentiment.

Chervachidze and Wheaton (2013) extend their analysis of risk premia with macroeconomic variables. The growth rate of debt relative to GDP incorporates information about liquidity, which significantly influences the cap rate. Duca and Ling (2015) examine the impact of the latest financial crisis on the commercial real estate market in the USA. Picking up from the work of Chervachidze and Wheaton (2013), they define the risk premium as the spread of the Baa corporate yield and the ten-year Treasury yield. By using this spread as a risk measure, they

stress the importance of linking investment market swings to the broader national economy, which will reflect back into the real estate market.

Shilling and Sing (2007) utilize the findings of Sivitanides et al. (2001) and Hendershott and MacGregor (2005a, 2005b), and extend their research on yields with a focus on the rationality of real estate investors and define rationality as the difference between the realized and the expected return on investment. According to the authors, unreasonable expectations do have a negative impact on returns and should, therefore, be considered in a modelling framework. Chichernea et al. (2008) show that geographical differences among the examined Metropolitan Statistical Areas (MSAs) influence real estate yields. The authors examine both the demand and supply side of the different local real estate markets and find that supply-side constraints have a stronger impact on cap rate variations than direct growth measures. In general, they establish that markets with higher liquidity and markets with more stringent supply constraints experience lower yield levels.

3.3 THEORY

Given the fact that the literature review (chapter 3.2) has revealed that sentiment indices are widely excluded from yield models, with the exception of Clayton et al. (2009), it is worth elaborating on the expected behaviour of the sentiment indicators in the yield models. As shown in various studies, such as Tsolacos (2012), the European commercial real estate market is subject to sentiment. I am therefore confident that an irrational or human element within the yield model will enable us to improve the model.

In addition, the literature review (chapter 2.1.2) has shown that the distinction between direct and indirect sentiment proxies has been applied in equity and real estate markets. Since this study covers 24 European countries, data availability plays an important role, especially when it comes to direct real estate specific sentiment indicators. For the British market, the Royal Institute of Chartered Surveyors (RICS) publishes a property survey, where RICS members are asked about their opinions on future developments in the real estate market. However, the majority of the remaining European countries do not offer an equivalent.

For this reason, we have to employ indirect sentiment proxies to mirror market perceptions. Yet the quantification of sentiment based on indirect sentiment proxies remains a crucial process. Following the basic idea of Baker and Wurgler (2006, 2007) and its application by Ling

et al. (2014) on the real estate market, that each imperfect sentiment proxy, at least to a certain degree, carries some pure sentiment, I am confident of extracting sentiment from indirect measures.

Many Eastern European countries do not offer data to the same extent as some Western European countries. This makes it difficult to follow the literature when it comes to the selection of sentiment proxies (Lee et al., 1991; Clayton et al., 2009; or Ling et al., 2014).

Ling et al. (2014), for instance, used one survey-based measure from the Real Estate Research Corporation (RERC) and eight different indirect sentiment proxies (the REIT stock price premium to the Net Asset Value (NAV), the percentage of properties sold each quarter from the NCREIF index, the REIT share turnover, the number of REIT Initial Public Offerings (IPOs), the average first-day returns, the share of net REIT equity issues relative to total net REIT debt issues, the net commercial mortgage flow as a percentage of GDP, and the net capital flows to dedicated REIT mutual funds). These proxies share a relative focus on the REIT market in the USA. More mature Western European countries such as the U.K., Germany or France are able to show a healthy REIT market. However, Eastern European countries do not have similar markets and especially not at the same depth.

In the methodology section, I will explain the intention and construction of the four different sentiment indicators. However, two things should be pointed out at this stage. First, I assume that the measured sentiment should have a negative impact on property yields. Since it is the intention of this study to capture investor sentiment, a negative relationship between yields and sentiment seems logical. The higher the sentiment the larger is the downward effect on the yields. This intuition can be explained by the assumption, that investors have an interest in rising property prices, which is associated with lower yields. Again, the yield is defined as the NOI over the market price.

Second, I follow the overall belief that direct real estate markets, given short-selling constraints and limits to arbitrage, incorporate mispricing of their properties. Nevertheless, the literature review has left the impression that scholars in the real estate market, even though they emphasize that they measure the sentiment of investors, do not follow an entirely behavioural approach. Their definition of irrationality is, instead, based on the incompleteness of classical financial theories, which is caused by the real estate market structure. In Baker and Wurgler (2007), the sentiment is defined as the belief of investors about future cash flows and investment risk that is not justified by the facts at hand. This belief is easily quantified with direct sentiment measures, which are based on the opinions of market participants and incorporate

forward-looking elements (Freybote, 2016). Using indirect measures (e.g. REIT share turnover), on the other hand, the aggregated belief of investors should be equal to the unexplainable part. This is why orthogonalization in combination with a principal component analysis (PCA) should provide a good indication of the actual irrationality of market participants.

3.4 METHODOLOGY

In this section, I will outline the components of a standard yield model. Subsequently, I will discuss the construction of the four sentiment measures, namely a macroeconomic, two real estate specific (office and retail) and a Google Trends sentiment measure – these will enter the standard yield model.

3.4.1 YIELD MODEL

Critical components in the primary yield model are the risk-free rate, the expected rent, and the risk premium. Equation 3:1 presents the basic panel model for yields.

$$Yield_{(office\ or\ retail)r,t} = \beta_o + \beta_1 Risk\ Free\ Rate_{j,t} + \beta_2 Risk\ Premium_{j,t} + \beta_3 Expected\ rent_{j,t} + \beta_4 regional\ fixed\ effect_r + \varepsilon_{j,t} \quad \textit{Equation 3:1}$$

where j represents the country, t is time and r is the specific city region. The random error term ε_{jt} is an independent and identically distributed (i.i.d.) error that embodies other time series and cross-sectional effects.

The transaction-based prime yield for office and retail has been provided by DTZ. The property yield is a function of the net operating income from real estate assets and the market price. Using a transaction-based yield allows a better insight into the market. The yield should incorporate the current situation within the market. While contractual rents are usually fixed over longer periods, prices are influenced by the negotiation of two parties and various market factors. Among others, the expectations about the market development influence the price as

well. Therefore, the yield should be subject to sentiment swings and yield models should subsequently benefit from the consideration of sentiment measures. Possible measurement issues for countries with in-transparent markets could result in insufficient market data. Markets where it is uncommon to report transactions publicly, service agencies struggle to get a full market coverage. Published yields, on the other hand, should, therefore, not be taken as a general market yield, since they might not mirror the actual market development.[3]

Earlier I highlighted the importance given to *expected rents* in yield determination. Most scholars agree that the rent component should carry the expectations of landlords and investors (Sivitanidou and Sivitanides, 1999) as well as regional influences (Hendershott and MacGregor, 2005a). Of the effective methods for calculating expected rent (Hendershott and MacGregor, 2005a, 2005b; Chervachidze and Wheaton, 2013) I have chosen Hendershott and MacGregor's approach and construct the rent variable as a four-quarter moving average of the long-term deviation of the log of real rents. This allows us to consider the slow adjustment of the market, which is captured as the moving average.

As the *risk-free rate*, I use the ten-year government bond rate for each country. I follow the work of Devaney et al. (2016), who calculated the *risk premium* as the volatility of the equity market. This is constructed as an eight-quarter rolling standard deviation from the stock market return. I consider this method consistent across all countries as data availability problems for some countries exist. Other methods based on the Baa bond rating, for example, are unavailable since the data is not present for all countries. An alternative method could have been the spread between either the German Bund rate or the yield rate from the European Union as a reference point. However, I thought these methods might be unsuitable since some countries are not members of the EU and for the German market the risk-free rate would have been zero throughout. Using such a long period for the construction of the risk measure (eight quarters) allows capturing an entire economic cycle. Depending on the volatility of the equity market, one could draw conclusions about the risk appetite of investors as well as the pricing in the market.

3.4.2 SENTIMENT MEASURES

As pointed out earlier this first analysis covers 24 European countries. Unfortunately, not all countries offer a direct real estate sentiment measure. Therefore, the use of sentiment proxies

[3] It is unknow how the data has been collected by DTZ. The provided dataset mainly reveals DTZ itself as the source of the various yields.

is the only solution to cover all countries and to give an opinion about country-specific sentiment.

The quantification of sentiment, based on indirect sentiment proxies, remains a crucial process. This became apparent in the literature review that the method developed by Baker and Wurgler (2006, 2007), using orthogonalization for the extraction of sentiment, is widely established.

Following Baker and Wurgler (2007), sentiment is the belief of investors that investment risk is not justified by the facts at hand. This belief is easily quantified with direct sentiment measures, which are based on the opinions of market participants and incorporate forward-looking elements (Freybote, 2016). Using indirect measures (e.g. REIT share turnover), on the other hand, the sentiment is not identified immediately, and those indirect measures need to be separated into obvious and unexplainable parts. This is why orthogonalization in combination with a principal component analysis (PCA) should provide a good indication of the actual irrationality.

3.4.2.1 MACROECONOMIC SENTIMENT INDICATOR

With regards to the yield modelling process and the influence of the economy on the real estate market, I assume that macroeconomic sentiment proxies contain information about market sentiment. Therefore, the first sentiment indicator is based on pure macroeconomic factors. Similar to Ling et al. (2014) I combine two direct sentiment proxies and four indirect sentiment proxies.

The first direct sentiment proxy is the Economic Sentiment Indicator (ESI) also used by Tsolacos (2012). The ESI is published by the European Commission and is a composite indicator of five weighted sector-specific confidence surveys covering construction (5%), retail (5%), industrial (40%), services (30%) and consumer sectors (20%). The indicator provides a good signal of the economic developments across countries and the general economic sentiment.

The second direct proxy is the Business Climate Indicator (BCI) also published by the European Commission, which provides a timely composite indicator for the manufacturing sector in the Eurozone. This indicator is based on five opinions from an industry survey: production trends in recent months, order books, export order books, stocks, and production expectations. These questions aim to retrieve the forward-looking opinions of market participants.

It might be misleading to combine direct and indirect sentiment proxies in order to construct an overall macroeconomic sentiment measure. However, the two presented direct sentiment measures, do not measure the real estate markets solely. As stated above the ESI measure does only account 5% of its weight to the construction industry. The BCI on the other hand does look on the manufacturing sector mainly and ignores the real estate industry. However, both measures reveal a lot about the general market development. Therefore, a statistical modification of the two measures is recommended, since they can only be seen as "indirect" sentiment proxies for the real estate market.

The indirect sentiment proxies should closely reflect general sentiment in the economy and, for consistency, they should be available across all countries. Four indirect series are selected. The stock market is considered a good indicator of national economic conditions. Among others, Baker and Wurgler (2006, 2007), Tetlock (2007) and Kurov (2010) find that investor sentiment influences stock markets. For each of the 24 countries in this study, I use the quarterly stock market returns. The data is provided by Thomson Reuters Datastream.

Similar to the stock index, the government bond rate can be used as an indicator of national economic health. This indicator is less likely to change as sharply as stock market returns; however, the government bond provides information about several country-specific risks, such as inflation, interest rate risk and the state of public finances.

Consumer confidence has been at the centre of interest since Katona (1968). Markets and governments are interested in which direction consumer confidence is heading. Therefore, consumer confidence is identified as a suitable sentiment proxy. Consumer confidence data are taken from the Organisation for Economic Co-operation and Development (OECD). I assume that this indicator can pick up some developments from consumer behaviour, that will feed into the real estate market sentiment.

Credit rating is the fourth indirect measure. It can be seen as an indicator, showing how a country is valued based on a range of macroeconomic factors. The credit rating is likely to be one of the primary indicators foreign investors focus on before they make an investment decision. The credit rating figures are provided by Oxford Economics and range between 0 and 20, where 20 equals a AAA rating.

To derive a suitable sentiment indicator, I apply an orthogonalization process to both the direct and indirect sentiment proxies and try to remove known macroeconomic influences. The focus is set on the main factors, such as the change of GDP, the forecast change in GDP, the

interest rate, the logarithm of the consumer price indicator, the logarithm of consumer spending, the unemployment rate, as well as the percentage change of the industry production of the country ($c\_gdp, fc\_gdp, intr, logcpi, logcsp, unemp, indpropc$).

The process requires that each of the proxies is regressed against those factors (macroeconomic influences) without an intercept. The residuals of these six orthogonalization regressions (for two direct and four indirect sentiment measures) are taken to reflect the market instinct and the unexplained part within the different sentiment measures. After the known components have been removed (i.e. GDP and interest rate) the remainder should be a proxy of the "gut-feeling" of the market.

Following Baker and Wurgler (2006), the residuals are standardized and, due to the fact that some variables may react to changes in the sentiment more rapidly than others, it is recommended to use both the standardized variables and a lagged version of them in a PCA. I obtain the first principal component with the highest eigenvalue. I calculate the correlation between the factor loadings and the first stage index from the PCA. Factor loadings with a small correlation are removed from the final sentiment calculation. Finally, the correlation between the first stage index and the constructed sentiment indicator is measured, to clarify if there is any severe loss of information by removing the weaker factors. This combines the six proxies to the macroeconomic sentiment indicator.

3.4.2.2 REAL ESTATE SPECIFIC SENTIMENT INDICATORS

The second and third indicators are designed to approximate the commercial real estate specific sentiment. I assume that a sentiment indicator based on property-specific elements that are monitored by market participants will contain more market-specific information compared to a solely macroeconomic sentiment indicator. To obtain a sentiment proxy that covers most European countries, I make use of commercial total return series from *MSCI* - IPD. Total returns embody sentiment swings in the commercial property market. However, the use of this sentiment proxy leads to an overall reduction of the city regions in the sample by 13, since the return series is not published for all countries.

The real estate data which is used in this study has been provided by Cushman & Wakefield (formerly known as DTZ). Other property-specific factors, such as demand and supply, also affect sentiment as market participants base their views on demand and supply data. For offices,

Cushman & Wakefield provides data for rent, office supply, office availability, office take-up, office availability ratio and office new supply as well as the yield.

The service provider defines the various office specific factors as follows. The provided rent is the local headline rent. The variable does not consider any concessions and it can be assumed that the rent represents the actually paid square meter price.

Office supply is the area which is completed by developers. Cushman & Wakefield further considers second-hand supply, which is space that has become available by tenants moving to a new space.

Office availability is all marketed spaces, that is available to move into within the next six months. Space does not have to be vacant at the current stage.

According to the service provider, office take-up is measured by occupational transactions. Office spaces are considered to be those which are let or sold to an eventual occupier. Further new developments which are either pre-let or sold to an occupier, as well as purchases of freehold or long leaseholds, are considered in this category.

The office availability ratio is defined as office space currently available as a percentage of stock projected six months ahead (i.e. includes speculative completions during that period).

Office new supply is floor space that has become newly available within the market, including developments within the next six months and all units available from the second-hand market.

Since these are the observed factors, I follow the same process as described in the previous section and orthogonalize the IPD total return for offices against these factors to obtain the residuals. Since only one proxy is used, there is no need for a PCA to retrieve a standard sentiment component. In the end, I have standardized the residuals.

On the retail side, the dataset is limited. Besides the retail yield, which will be used as the dependent variable, only the headline rent is available. Again, the IPD total return for retail is then orthogonalized against the rent. I am aware that this results in a less informative sentiment indicator since I am unable to remove more obvious market factors from the chosen sentiment proxy.

Next, I have constructed another set of five indicators. They are mainly used for robustness checks with the intention of testing the methodology as well as testing if the chosen sentiment factors are superior in the way they are compiled.

The fourth indicator uses only a PCA on the six sentiment proxies. The idea behind this method is to check if the orthogonalization is needed to create a superior indicator. This second macroeconomic sentiment indicator will be tested against the other macroeconomic indicator.

The fifth sentiment indicator is used to check if the recommended use of the first stage index is suitable since it ignores the Kaiser Criterion in the PCA. The Kaiser Criterion states that all components with an eigenvalue above 1 should be included in the process.

Since the two property specific indicators have been generated without the use of a PCA, I have created a sixth indicator, which checks if a PCA of the two property sentiment indicators can produce a combined property sentiment indicator.

Following a similar intention, the seventh indicator adds the two property specific indicators to a single such indicator.

The last indicator which is based on the office- and property specific variables is constructed in a similar fashion as the retail-specific indicator. I have only orthogonalized the office prime rent from the IPD total return for offices.

3.4.2.3 SENTIMENT CONSTRUCTION

It is worth illustrating the sentiment construction process in more detail. I will, therefore, provide a step-by-step guide of how the sentiment indicator has been derived.

I will first give a short introduction to the process of PCA and orthogonalization.

3.4.2.3.1 PRINCIPAL COMPONENT SENTIMENT INDICATORS

PCA belongs to the class of factor models and is used when explanatory variables are closely related, as in this case, when it is assumed that the proxies share a common component. The model transforms k explanatory variables into k uncorrelated new variables. The new principal components are independent linear combinations of the original data. Assume that the original variables are symbolized by $x_1, x_2, \dots, x_k$ and the principal components are symbolized by $p_1, p_2, \dots, p_k$, then

$$\begin{aligned} p_1 &= \alpha_{11}x_1 + \alpha_{12}x_2 + \cdots + \alpha_{13}x_3 \\ p_2 &= \alpha_{21}x_1 + \alpha_{22}x_2 + \cdots + \alpha_{23}x_3 \\ p_k &= \alpha_{k1}x_1 + \alpha_{k2}x_2 + \cdots + \alpha_{k3}x_3 \end{aligned} \qquad \textit{Equation 3:2}$$

where α_{ij} are coefficients to be calculated, representing the coefficients on the jth explanatory variable in the ith principal component. These components are also known as factor loadings. Even though the theoretical approach suggests using all components with an eigenvalue above one, the Baker and Wurgler (2006) approach uses only the first component. This component usually incorporates the largest explanatory proportion. The estimated regression based on the first principal component would be

$$y_t = y_0 + y_1 p_{1t} + \cdots + y_r p_{rt} + u_t \qquad \textit{Equation 3:3}$$

here y_t is the dependent variable, and y_0 to y_r present the estimated coefficients also known as β. p_{1t} states the first principal component for the first variable. Depending on how many independent variables are used r variables are added. u_t states the error term.

Due to the fact that some variables may react to changes in the sentiment faster than others, it is recommended to use both the standardized variable and a lagged version of them. Comparing the results of those loadings it has been decided to use those ones which have a higher correlation with the first stage index. Compared to the original OLS estimates the principal component estimates will be biased, but still will be more efficient since redundant information has been removed.[4]

3.4.2.3.2 ORTHOGONALIZATION

The theoretical and methodological approach is based on the Gram-Schmidt Algorithm and has been used by Baker and Wurgler (2006) and Ling et al. (2013). Suppose a univariate model with no intercept is given

$$Y = X\beta + \varepsilon \qquad \textit{Equation 3:4}$$

with the least squares and the residuals given by

$$\hat{\beta} = \frac{\sum_1^N x_i y_i}{\sum_1^N x_i^2} \qquad \textit{Equation 3:5}$$

$$r_i = y_i - x_i \hat{\beta} \qquad \textit{Equation 3:6}$$

In vector notation, we let $y = (y_1, \ldots, y_N)^T$, $x = (x_1, \ldots, x_N)^T$ and define the inner product between x and y_1

$$(x, y) = \sum_{i=1}^{N} x_i y_i \qquad \textit{Equation 3:7}$$

[4] See Brooks, 2014, p. 170.

$$= x^T y$$ *Equation 3:8*

This leads to,

$$\hat{\beta} = \frac{(x, y)}{(x, x)}$$ *Equation 3:9*

$$r = y - x\hat{\beta}$$ *Equation 3:10*

This is the base for a multilinear regression, where the inputs $x_1, x_2, \dots, x_p$ are orthogonal; $(x_j, x_k) = 0$ for all $j \neq k$. It can be shown that the multiple least squares estimates are equal to the univariate estimates. They are orthogonal and do not have any impact on each other's parameters in the models.

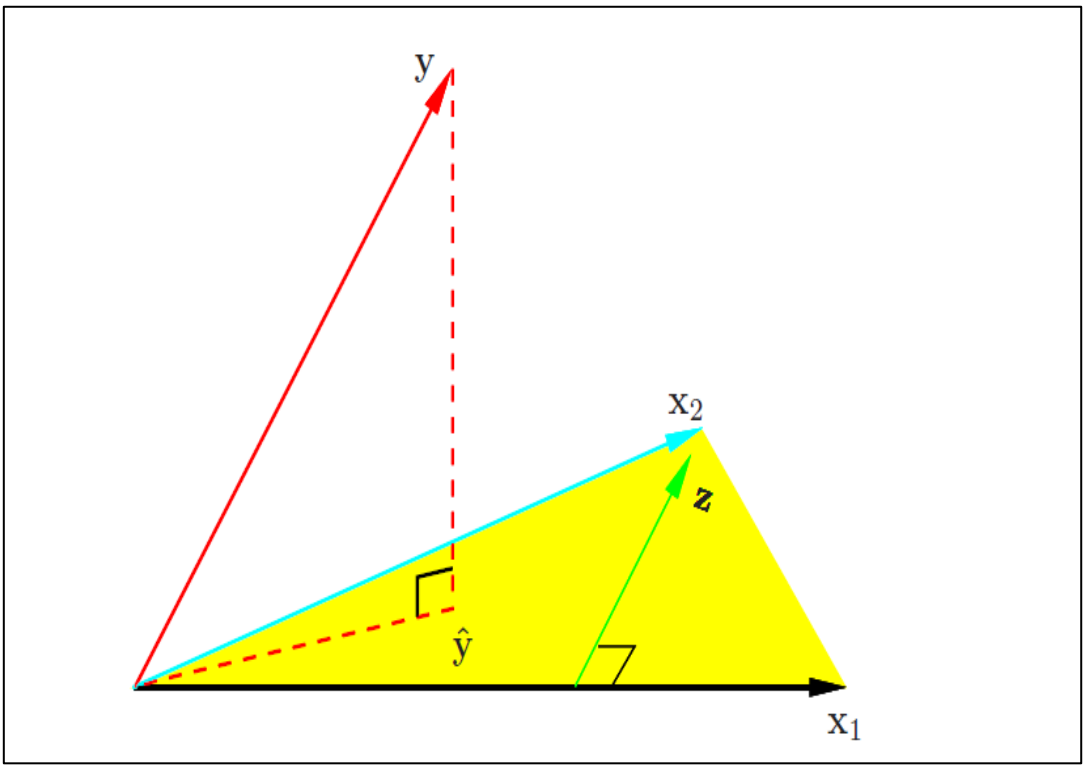
$$\widehat{\beta_1} = \frac{(x - \bar{x}1, y)}{(x - \bar{x}1, x - \bar{x}1)}$$ *Equation 3:11*

where $\bar{x} = \sum_i x_i / N$, and $1 = x_0$, the vector of N ones. Equation 3:11 is the result of two steps: (1) regress x on 1 to produce the residuals $z = x - \bar{x}1$; and (2) regress y on the residuals z to give the coefficient $\widehat{\beta_1}$.

This approach means a simple regression of b on a with no intercept, and produces coefficients and residual vectors. b is orthogonalized with respect to a. This process does not change the parameters but produces an orthogonal basis for representing it. The general idea is to extract a latent component which is incorporated in one of those elements.

Figure 3:1 illustrates the Gram-Schmidt Algorithm. Vector x_2 is regressed on x_1 and produces the residual vector z. Regressing y on z will give the coefficient for the multiple regression of x_2.

Figure 3:1 - Gram-Schmidt Algorithm



Note 3.1 - Source: Hastie et al. (2008), p. 54

3.4.2.3.3 MACROECONOMIC SENTIMENT

The leading macroeconomic indicator is constructed with the orthogonalization and PCA process. In a first step, I have checked for any apparent correlations between the sentiment proxies and the macroeconomic factors. Table 3:2 illustrates the correlation coefficients. It can be seen that most of the correlations are weak to moderate. Only the combination of the interest rate and the 10-year government bond rate shows a strong positive correlation of 0.798. This is, however, reasonable since both series are interlinked.

Table 3:2 - Correlation (macroeconomic sentiment)

	Economic sentiment indicator	Change of the stock market	Change of the consumer confidence	Credit rating	10-year government bond rate	Business climate indicator
Change of GDP	0.126	0.187	0.083	-0.027	-0.058	0.190
Forecasted change of GDP	0.246	0.060	0.238	-0.185	0.290	0.383
Log of consumer price index	-0.068	-0.024	0.110	-0.203	0.248	-0.012
Interest rate	0.127	-0.059	0.020	-0.402	0.798	0.129
Log of consumer spending	0.161	-0.028	-0.224	0.441	-0.156	0.062
Unemployment rate	-0.105	0.076	-0.180	-0.303	0.082	-0.129
Percentage change of the industry production of the country	0.273	0.417	0.195	-0.127	0.049	0.443

Note 3.2: The table illustrates the correlation between the macroeconomic factors and the sentiment proxies.

Starting with the orthogonalization process, the macroeconomic factors will be regressed against the sentiment proxies. The regression is run without an intercept. The residuals which are obtained from these six regressions are assumed to resemble the unexplained part. Table 3:3 provides the regression results. Since the process is not targeted on the provided statistics of the regression but on the residuals produced by this process, I will not comment on the results.

Table 3:3 - Regression results of the orthogonalization process (macroeconomic sentiment)

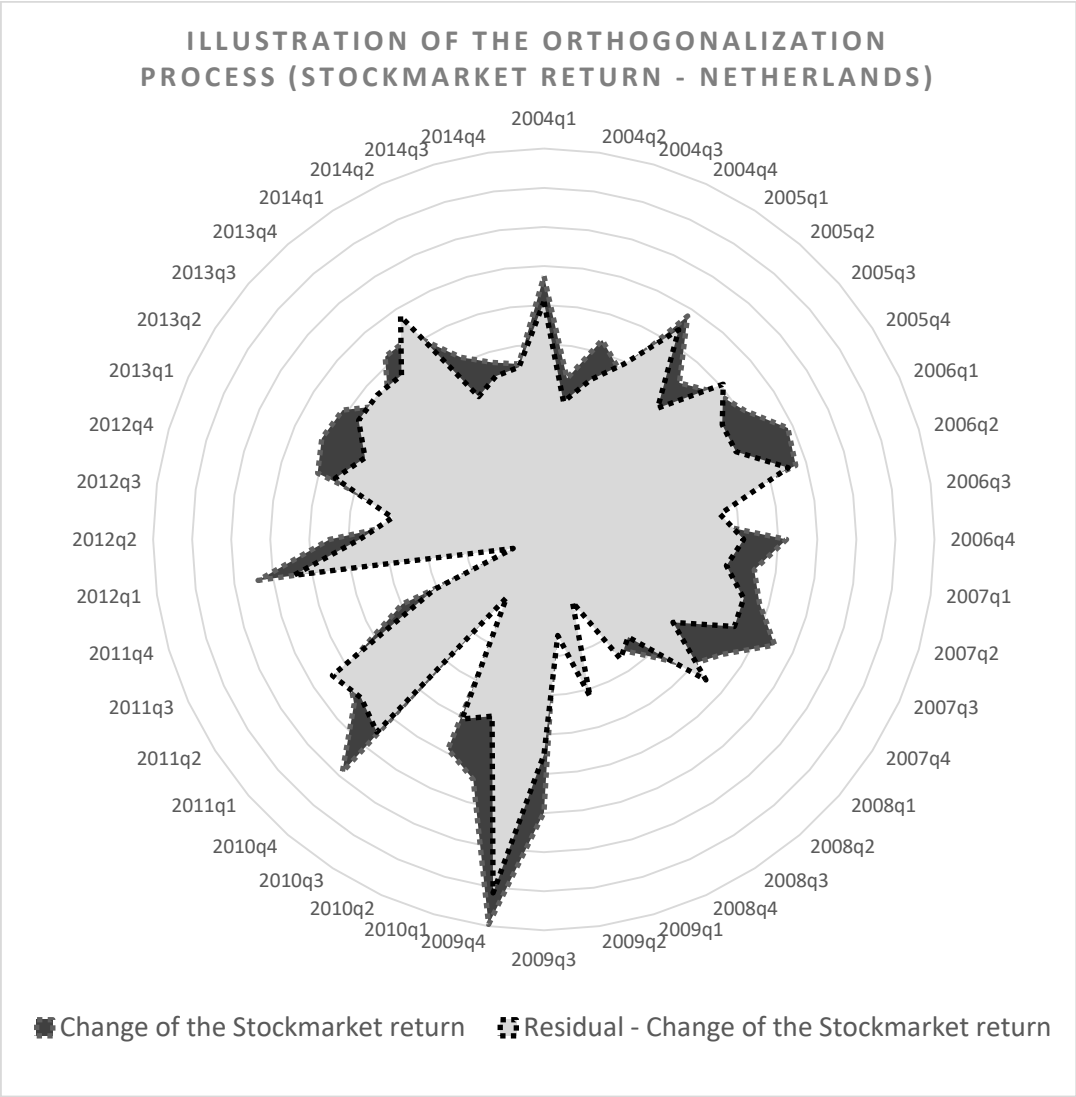
VARIABLES	LABELS	Economic sentiment indicator	Change of the stock market	Change of the consumer confidence	Credit rating	10-year government bond rate	Business climate indicator
c_gdp	Change of GDP	30.103***	17.942***	28.499***	1.05	-2.104***	9.144***
		[8.674]	[6.309]	[9.782]	[0.832]	[0.707]	[3.207]
fc_gdp	Forecasted change of GDP	654.059***	-41.393	610.482***	15.962	0.231	363.666***
		[113.867]	[43.520]	[110.230]	[24.207]	[11.666]	[86.906]
logcpi	Log of Consumer Price Index	1.611**	0.018	0.438	0.210**	0.069**	2.535***
		[0.737]	[0.057]	[0.930]	[0.083]	[0.033]	[0.336]
Intr	Interest rate	0.857*	-0.048	1.373	-0.536***	0.606***	0.317
		[0.493]	[0.090]	[1.143]	[0.082]	[0.025]	[0.307]
logcsp	Log of consumer spending	7.254***	0.107***	-0.678	1.726***	0.106***	7.027***
		[0.272]	[0.029]	[0.579]	[0.037]	[0.020]	[0.239]
unemp	Unemployment rate	0.458	0.121***	-0.752	-0.175***	0.127***	1.009***
		[0.357]	[0.036]	[0.478]	[0.046]	[0.026]	[0.328]
indpropc	Industry production	1.738***	1.445***	1.538***	-0.080***	-0.006	0.267***
		[0.197]	[0.156]	[0.268]	[0.021]	[0.011]	[0.092]
Observations		3,212	3,220	3,364	3,356	3,279	3,301
R-squared		0.972	0.171	0.143	0.979	0.93	0.992
Adjusted R-squared		0.972	0.17	0.141	0.979	0.93	0.992
F-statistics		4662	70.78	13.27	2369	1184	4937
Degrees of freedom		75	75	79	79	78	79
Number of clusters		76	76	80	80	79	80

Robust standard errors in brackets; *** p<0.01, ** p<0.05, * p<0.1

Note 3.3: The table illustrates the regression results of the orthogonalization process. In each of the six regressions, the constant is omitted.

Figure 3:2 illustrates the process in a graphical way. It can be seen that the residual (light shaded area) is for many quarters smaller in magnitude than the original variable (dark shaded area). This difference was caused by the observable factors.

Figure 3:2 - Orthogonalization process



Note 3.4: The spider-chart illustrates the process of orthogonalization. The change of the stock market return has been orthogonalized against the various macroeconomic factors. This has changed the magnitude of the variable for each period.

The obtained residuals will be now standardized with a mean of 0 and standard deviation of 1. Further, a lagged version of each variable is created. As pointed out earlier this should control for the case when some variables react earlier than others.

The lagged and unlagged variables now enter the PCA. Table 3:4 shows the results of the PCA. The applied methodology suggests the usage of the first component with the highest eigenvalue (3.293). The first component has a proportion of nearly 30% and therefore carries the most substantial weight.

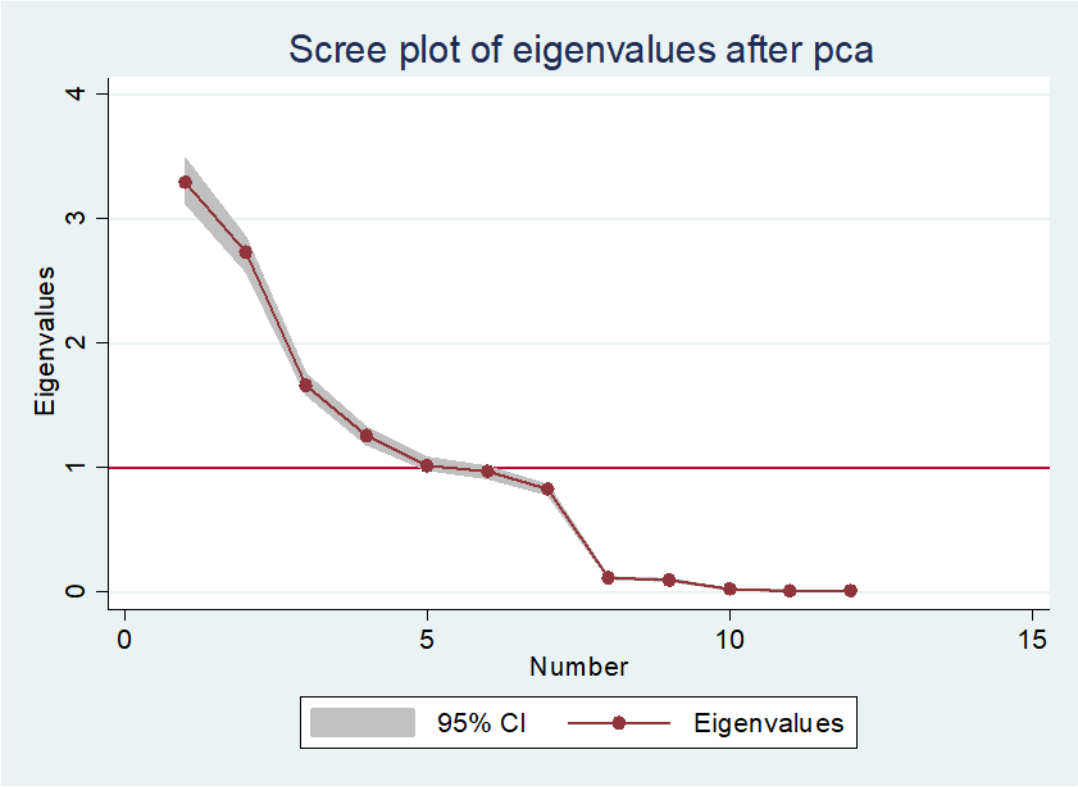
Table 3:4 - Principal component analysis (macroeconomic sentiment)

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	3.293	0.564	0.274	0.274
Comp2	2.729	1.066	0.227	0.502
Comp3	1.662	0.407	0.139	0.640
Comp4	1.255	0.238	0.105	0.745
Comp5	1.017	0.047	0.085	0.830
Comp6	0.970	0.146	0.081	0.911
Comp7	0.824	0.712	0.069	0.979
Comp8	0.112	0.015	0.009	0.989
Comp9	0.096	0.074	0.008	0.997
Comp10	0.022	0.011	0.002	0.998
Comp11	0.011	0.003	0.001	0.999
Comp12	0.008	.	0.001	1.000

Note 3.5: The table illustrates the result of the PCA. It can be seen that a total of 10 components have been found. Each component carries a certain proportion of explanatory power. Both the proportion value as well as the Eigenvalue decrease with each additional component. Therefore, the largest Eigenvalue is always assigned to the first component.

Figure 3:3 shows the corresponding scree plot and how the eigenvalues decrease with every new component.

Figure 3:3 - Scree plot of eigenvalues after PCA (macroeconomic sentiment)



Note 3.6: The scree plot illustrates the decrease of the Eigenvalues. Eigenvalues below 1 are assumed to be weak.

Each component from the PCA is the sum of the 12 proxy residuals which have entered the process. However, not all 12 residuals should build the sentiment, since they are mostly a twofold part of the component. Therefore, those components will be removed from the final sentiment construction, which have a smaller correlation (see Table 3:5 bold variables) with the first component.

Table 3:5 - Correlation between the residuals and the first component

LABELS	Correlation	Scoring coefficient component 1
First component	1.000	
The standardized residual of the ESI	0.522	0.288
The standardized residual of the ESI (1 lag)	0.538	0.297
The standardized residual of the change of the stock market return	0.024	0.013
The standardized residual of the change of the stock market return (1 lag)	0.054	0.030
The standardized residual of the change of consumer confidence	0.263	0.145
The standardized residual of the change of consumer confidence (1 lag)	0.275	0.152
The standardized residual of the credit rating	0.735	0.405
The standardized residual of the credit rating (1 lag)	0.721	0.398
The standardized residual of the 10-year government bond rate	-0.326	-0.180
The standardized residual of the 10-year government bond rate (1 lag)	-0.321	-0.177
The standardized residual of the BCI	0.811	0.447
The standardized residual of the BCI (1 lag)	0.809	0.446

Note 3.7: The table illustrates the correlation between the individual residuals and the first component. This analysis is performed to estimate which of the two residual variables should be used for the sentiment construction. According to the applied methodology, the residual variable with the highest (positive or negative correlation) enters the sentiment construction process. Bold variables will be ignored during the indicator construction.

Each selected residual variable will then be multiplied by its corresponding scoring coefficient from the PCA. All six sentiment proxies will then be aggregated to the macroeconomic sentiment indicator.

The last recommended test is another correlation analysis between the first component and the constructed sentiment indicator. The correlation should be reasonably high, which suggests that the removal of the remaining six factors has not removed much of the explanatory power. The correlation between the sentiment indicator and the first component is 0.994.

3.4.2.3.4 MACROECONOMIC SENTIMENT: KAISER CRITERION AND PCA ONLY

The other two mentioned macroeconomic indicators have been developed for robustness checks only. Both try to question the proposed method of Baker and Wurgler (2006).

Regarding the PCA, different approaches are discussed in academia. The proposed method focuses on the first principal component, which has the highest explanatory power. Nevertheless, academia uses a range of different methods to decide how many components should be included. Among others, the two primary methods are the Kaiser Criterion and the Scree Test. The Kaiser Criterion suggests using all components with an eigenvector above one. In the above-presented construction that would have meant that in total five components (Figure 3:3) should have been used. The difference to this construction lies in the fact that

virtually five sentiment indicators, based on the five principal components, have to be constructed. Therefore, one more step is required, which will combine the five indicators into one. I will use the corresponding weights of each component and multiply them by the indicator and aggregate the five at the end. The corresponding tables for the construction have been provided in the Appendix (Table 8:1 to Table 8:4).

The third indicator is trying to question whether the orthogonalization process is needed when the PCA is already looking for a component that is part of all sentiment proxies. As before the corresponding tables and graphs have been included in the Appendix (Table 8:5 and Table 8:6).

3.4.2.3.5 OFFICE SPECIFIC SENTIMENT

Since only one sentiment proxy has been used, the process of the PCA is obsolete. The six observable office factors will be orthogonalized from the sentiment proxy. For the main office sentiment indicator Table 3:6 provides the correlation coefficients among the sentiment proxy and the observable factors. The correlations range between weak and strong, with the highest correlation for the log of office availability and log of office supply (0.863).

Table 3:6 - Correlation between the IPD total return index and the six office factors

	IPD total return (offices)	Log of office rent	Log of office supply	Log of office availability	Office availability ratio	Log of office take-up	Log of office new supply
IPD total return (offices)	1.000						
Log of office rent	0.455	1.000					
Log of office supply	-0.253	0.068	1.000				
Log of office availability	-0.207	-0.043	0.863	1.000			
Office availability ratio	-0.009	-0.240	-0.057	0.424	1.000		
Log of office take-up	-0.316	0.134	0.564	0.528	0.045	1.000	
Log of office new supply	-0.266	-0.043	0.395	0.431	0.145	0.577	1.000

Note 3.8: The table illustrates the correlation between the sentiment proxy (IPD office total return) and the observable office factors.

Again, the sentiment proxy is regressed against the observable factors without an intercept. Table 3:7 provides the regression results for the pooled OLS for the panel dataset.

Table 3:7 - Orthogonalization process (office sentiment)

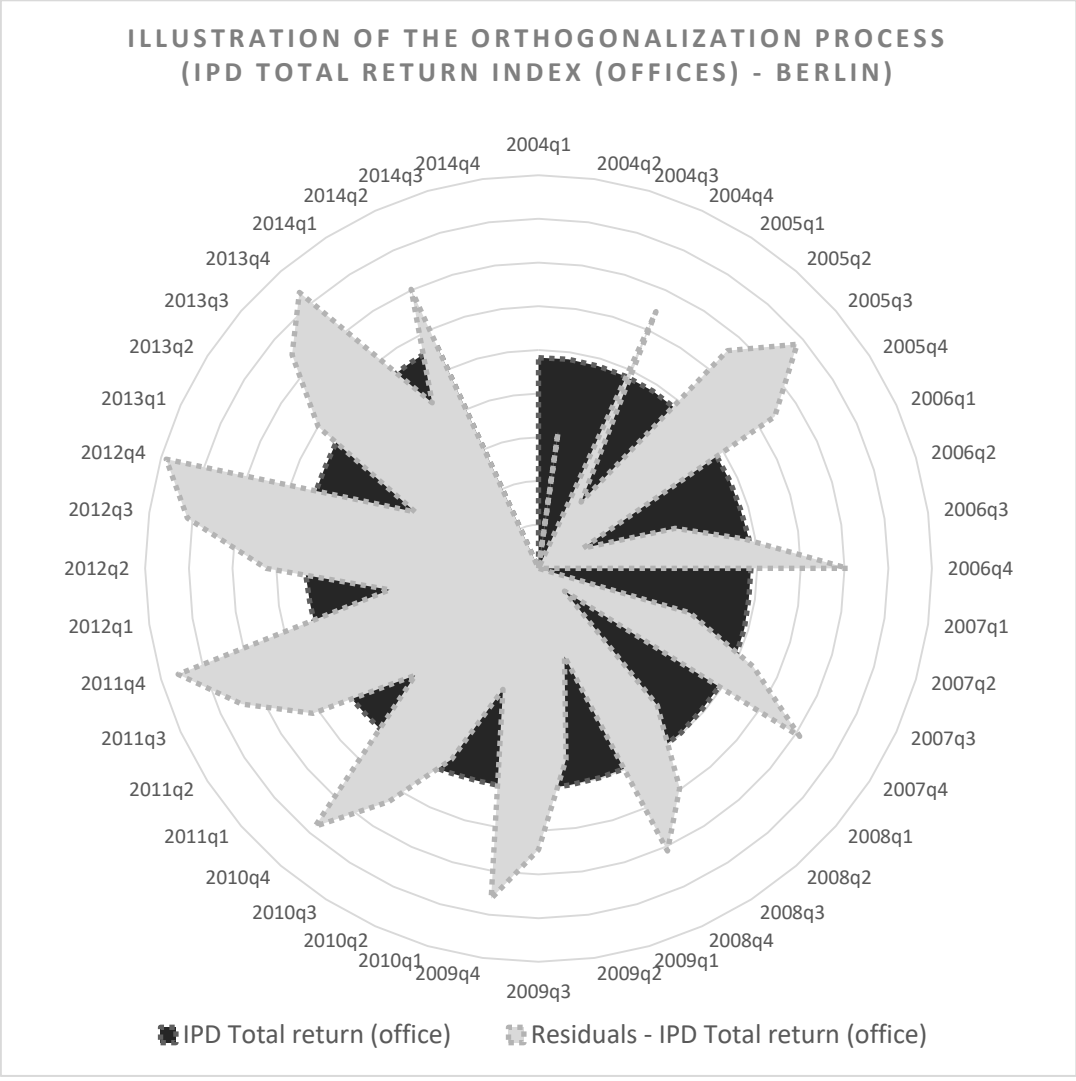
Variables	Labels	IPD - total return index (office)
Logofr	Log of office rent	552.614***
		[89.439]
Logofs	Log of office supply	6.626
		[122.860]
Logofa	Log of office availability	-3.927
		[143.050]
Ofar	Office availability ratio	20.725
		[17.243]
Logoftu	Log of office take-up	-146.514***
		[43.288]
Logofns	Log of office new supply	-13.046
		[18.478]
Observations		1,505
R-squared		0.563
adjusted R-squared		0.561
F-statistics		11.64
Degrees of freedom		58
Number of clusters		59

Robust standard errors in brackets; *** p<0.01, ** p<0.05, * p<0.1

Note 3.9: The table illustrates the regression results for the orthogonalization process for the office sentiment. As suggested by the methodology, the constant is omitted in the regression. Only two variables (the Log of office rent and the log of office take up) remain highly significant.

Figure 3:4 illustrates the process for Berlin. It can be seen that the process has not worked as it has before for the change of the stock market return for the Netherlands (Figure 3:2). The residual for the sentiment proxy is not smaller in most of the quarters. This indicates that the observable factors might not be as suitable as I had assumed before. However, in the absence of other property specific variables, I will proceed with the constructed sentiment variable. The presented result is unique for Berlin, since the independent variables are linked to the city-region level.

Figure 3:4 - Orthogonalization process: IPD total return index (offices) for Berlin



Note 3.10: The spider chart illustrates the difference between the IPD total return index for offices in Berlin and the residual from the orthogonalization process.

In the last step, the residuals have been standardized with a mean of 0 and a standard deviation of 1.

A second office-specific sentiment indicator has been developed. Since the retail-specific sentiment indicator (see below) can only rely on the headline rent, I have orthogonalized the headline rent office as well against the office sentiment proxy (Table 8:7). This should make the two indicators more comparable to each other.

3.4.2.3.6 RETAIL SPECIFIC SENTIMENT

As pointed out before, the dataset, unfortunately, does not offer more than one variable for retail. Therefore, the construction of the retail-specific sentiment indicator relies solely on the orthogonalization of the headline rent against the IPD total return index for retail.

The headline rent and the sentiment proxy have a positive moderate correlation of 0.486. Table 3:8 illustrates the orthogonalization process of the retail-specific sentiment indicator. The obtained residual is then standardized with a mean of 0 and a standard deviation of 1.

Table 3:8 - Orthogonalization process (retail sentiment)

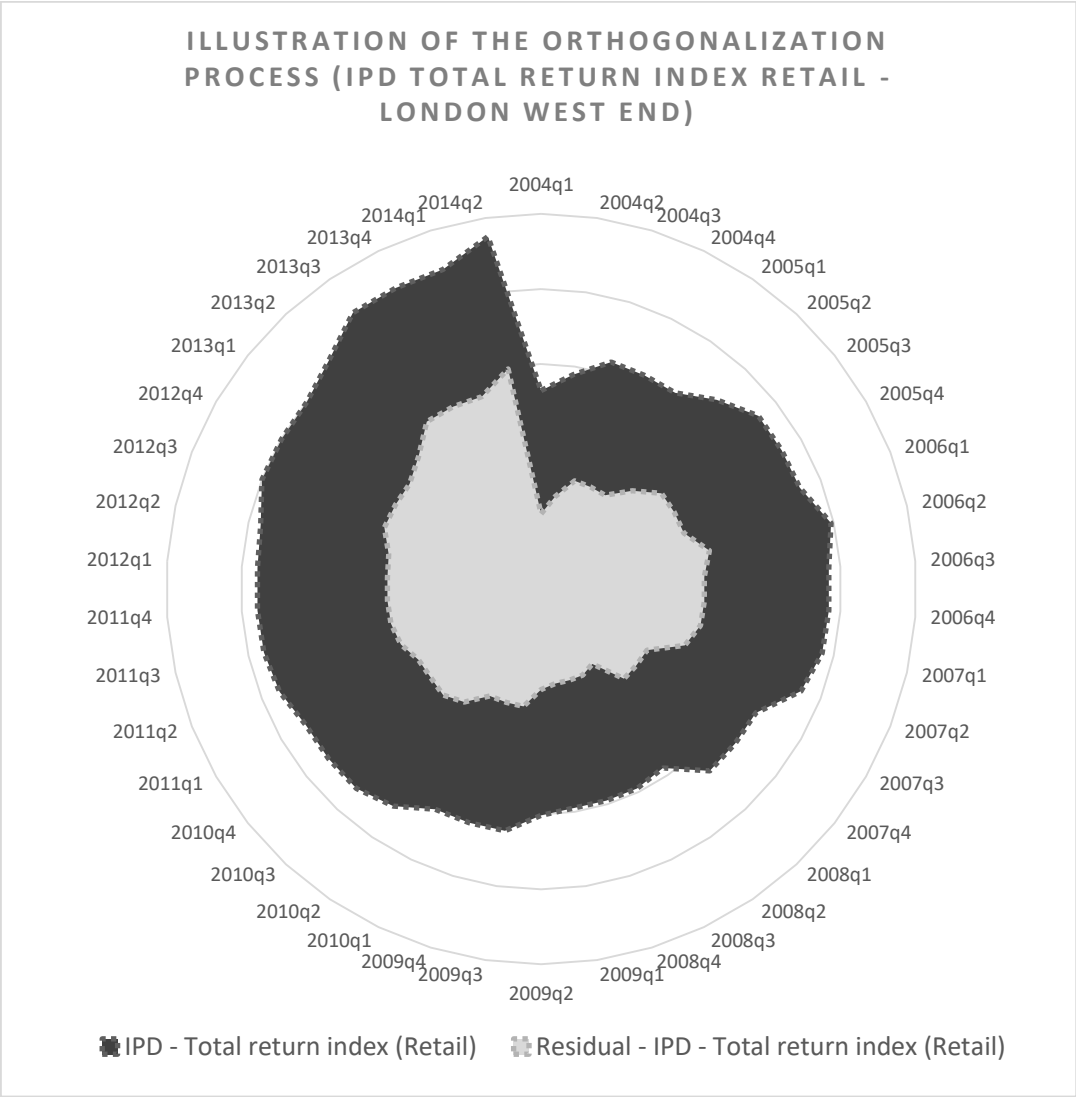
Variables	Labels	IPD: total return index (retail)
logretr	Log of retail rent	123.525***
		[21.509]
Observations		1,690
R-squared		0.465
Adjusted R-squared		0.464
F-statistics		32.98
Degrees of freedom		46
Number of clusters		47

Robust standard errors in brackets; *** p<0.01, ** p<0.05, * p<0.1

Note 3.11: The table illustrates the rather simple orthogonalization process for the retail measure.

Different to the previously presented result, Figure 3:5 shows that the orthogonalization process has produced sufficient results for the London West End market.

Figure 3:5 - Orthogonalization process: IPD total return index (retail) for London West End



Note 3.12: The figure illustrates the result of the orthogonalization of the IPD total return retail index for London West End. Different to the previous orthogonalization example (Figure 3:4), here the process has obviously reduced the magnitude of the dependent variable.

3.4.2.3.7 PROPERTY SPECIFIC SENTIMENT

The previous sentiment indicators are based on the two property specific indicators for office and retail. It is my intention to generate a composite property indicator which is based on the sentiment for both shares of the market. Both newly constructed indicators will be used for robustness checks.

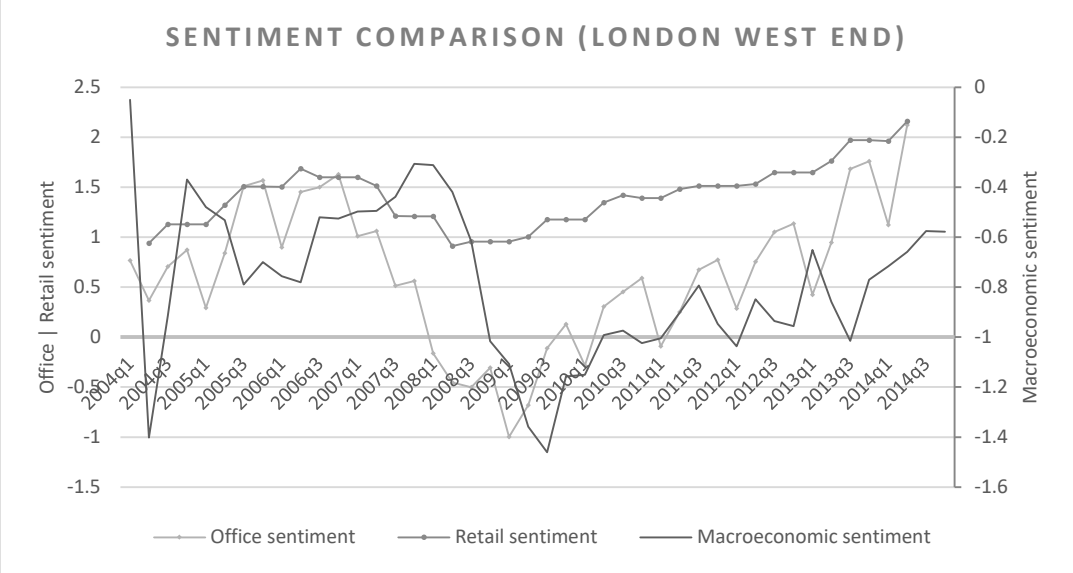
The first property specific indicator is based on a PCA. Here the office and the retail sentiment index (please see 3.4.2.3.5 & 3.4.2.3.6), as well as a lagged version of each indicator,

enter the PCA. Again, a composite index is constructed based on the correlation between the individual shares of the four variables and the first component (see Table 8:8 to Table 8:10).

The second indicator attempts a similar approach, where both primary indices are simply averaged. This should also provide a property market specific indication of the sentiment.

Figure 3:6 illustrates the three main sentiment indicators for the London West End market. While the two property specific indicators increase after the financial crisis, the macroeconomic indicator remains more or less stable with a slight trend upwards.

Figure 3:6 - Sentiment comparison for the London West End market



Note 3.13: The figure illustrates the three different sentiment indicators. It can be seen that the three-sentiment series show different developments. The retail series has the highest values. This is probably caused by the low number of observable factors which been removed in the orthogonalization process. The office sentiment indicator shows a rather cyclical development with a clear decrease over the cause of the financial crisis. The macroeconomic indicator, on the other hand, has the lowest values and shows a steady development, after the financial crisis.

3.4.2.3.8 GOOGLE TRENDS

The last sentiment indicator utilizes online search volume data. Studies such as Dietzel et al. (2014) show that online search volume data are able to give information about the thoughts of millions of people and their intentions.

Probably the majority of online searches are motivated by information collection. However, a proportion could also be triggered by “hot topics” within the market. In that scenario, these searches would not entirely reveal the actual interest in the search term. For the remaining

cases, where the search is performed to collect information, I assume that a later related action can be expected.

Using the Google categorization should filter out reactional searches. The searches related to a specific category, such as “property”, are only counted by the Google algorithm in this category if a series of property-related searches are performed. Similar to other studies, the analysis follows the belief that the volume of online searches within the specific category reflects the sentiment of the market and represents a suitable way of measuring the mood.

It remains unclear how professionals interact with the search engine. Some investors might have an in-house research department or rely on a network or their personal experiences. Given this, the contribution to the literature using Google data is twofold. First, a European-wide analysis of the commercial real estate market is performed. Europe is characterized by a variety of different national languages which makes a translation of the search terms necessary. Second, unlike Dietzel et al. (2014), this study does not solely rely on the broad search volume index (SVI), which is an aggregation of all category-specific (property) searches. The broad SVI incorporates other searches regarding the housing market and is therefore assumed to carry noise.

The constructed Google Trends index uses a set of 90 specific search words (Table 8:15) for each region within the dataset. These search words are partly focused on the office and retail property category, and partly focused on the market players, such as service agencies and banks. The intention is and this addresses the earlier criticism that institutional investors might not search online for an office property but will search for a telephone number or a market report from a service agency, which could result in an actual transaction. Therefore, this method is assumed to be able to capture these motivations in a more directed way.

Table 3:9 summarizes the different sentiment indicators with the acronym, their method and their summary of statistics.

Table 3:9 - Summary of statistics

Variable	Label	Method	Obs	Mean	Std. deviation	Min	Max
macroecono~t	Macroeconomic sentiment	Orthogonalization & PCA	2,863	-0.078	0.809	-2.468	2.926
me_sentime~c	Macroeconomic sentiment (Kaiser criterion)	Orthogonalization & PCA	2,858	-0.037	0.349	-1.002	1.610
pca_macroe~t	Macroeconomic sentiment (PCA)	PCA	3,010	0.000	1.334	-8.616	4.107
office_sen~t	Office sentiment	Orthogonalization	1,505	0.000	1.000	-2.283	3.474
retail_sen~t	Retail sentiment	Orthogonalization	1,690	0.000	1.000	-1.235	2.422
office_sen~2	Office sentiment (II)	Orthogonalization	2,519	0.000	1.000	-0.991	2.966
pca_proper~t	Property sentiment (I)	PCA	948	0.071	0.871	-1.367	2.819
property_s~t	Property sentiment (II)	Aggregation of the office and retail sentiment measure	3,520	0.000	0.560	-1.366	2.925
ZGT	Google Trends	Search volume analysis	3,300	0.000	1.000	-1.933	3.543

Note 3.14: The table above illustrates the summary of statistics for the eight constructed sentiment indicators. While the statistical values of the different sentiment measures are more or less similar, with the exception of the Macroeconomic sentiment measure constructed by PCA, the number of observations differ. The reason for these variances lies in the underlying difference in the methods and in the data availability. Not all sentiment proxies and not all macroeconomic/ real estate variables, have been available for all countries at all times. I refer to the descriptive statistics of the various variables used in this chapter (Table 8:13 and Table 8:14). The overview should provide enough insight, in where the data issues lie.

3.4.3 EMPIRICAL MODELS

The yield models, which are presented in the following, are based on a feasible generalized least squares approach. Test runs have revealed that common use of panel data quantification methods in form of random effects and fixed effects models lead to model specification issues. This method offers some benefits for the handling of panel data. Estimations are possible in the presence of AR (1) autocorrelation within panels and cross-sectional correlation and heteroskedasticity across panels. A vector autoregressive model (VAR) could have been used as well, in order to capture the linear interdependencies among the variables. The chosen method, however, does deal with missing observations and does produce reasonable results. Compared to a VAR model, the feasible generalized least squares approach seems less established and does still lack agreed guidance for a range of standard tests. Therefore, some benefits of the more established approach are missing. This issue is addressed at a later stage of this thesis again, and future research will consider an alternative modelling approach.

For each property type, a total of four yield models is estimated. Equation 3:1 is the base model, and it is estimated with no sentiment on the right-hand side for offices and retail. Equation 3:12 and Equation 3:13 augment the base model with the inclusion of (i) macroeconomic sentiment proxies, (ii) real estate market proxies or (iii) the Google Search Volume indicator. Equation 3:12 is the office equation and Equation 3:13 is the empirical framework for the retail sector.

$$ofy_{r,t} = \beta_o + \beta_1\, rf_{c,t} + \beta_2\, rprem_{c,t} + \beta_3\, ofr4qma_{r,t} + \beta_4\, regional\ fixed\ effect_r + \beta_5\, sent_{c,r,t-x} + \varepsilon_{j,t} \quad \textit{Equation 3:12}$$

where

$(logofy_{r,t})$ is the logarithm of the office yield specific for region (r) at time (t)

$(rf_{c,t})$ is the risk-free rate at country (c) at time (t)

$(rprem_{c,t})$ is the risk premium for country (c) at time (t)

$(r4qma_{r,t})$ is the deviation of real office rent from a four-quarter moving average in the city regions (r) at time (t)

$(regional\ fixed\ effect_r)$ represents regional fixed effects

$(sent_{c,r,t-x})$ represents one of the three different sentiment indicators: macroeconomic, office and online search volume sentiment.

$$rety_{r,t} = \beta_o + \beta_1\, rf_{c,t} + \beta_2\, rprem_{c,t} + \beta_3\, retr4qma_{r,t} + \beta_4\, regional\ fixed\ effect_r + \beta_5\, sent_{c,r,t-x} + \varepsilon_{j,t} \quad \textit{Equation 3:13}$$

where

$(logrety_{r,t})$ is the logarithm of the retail yield specific for regions (r) at time (t)

and different to above

$(retr4qma_{r,t})$ is the deviation of real retail rent from a four-quarter moving average in the city regions (r) at time (t)

$(sent_{c,r,t-x})$ represents one of the three different sentiment indicators: macroeconomic, retail and online search volume sentiment.

The remaining variables do not change compared to Equation 3:12. The model components, their source and the expected sign, are given in the Appendix Table 8:11.

3.5 DATA DESCRIPTION

This chapter analyses the European commercial real estate market from 2004q1 until 2014q4 (44 quarters), for 80 different regions spread out over 24 countries. The majority of countries are located in Europe, with the exception of Russia and Turkey. Some regions match entire cities. Other cities such as London or Paris are present multiple times in the dataset since some regions are specific economic regions, such as the Central Business District (CBD).

The dataset consists of real estate data for the office and retail markets and a range of macroeconomic variables. Cushman & Wakefield provided the real estate data. The macroeconomic data was collected via Thomson Reuters DataStream, the OECD, the International Monetary Fund (IMF) and through the European Commission. A panel dataset with 3,520 possible observations is constructed.

Some variables have missing observations. On the real estate side, the data is much more consistent for Western European countries than for Eastern European countries. The real estate variables include, among others, rents and yield values. For office, further take-up, stock, new supply, availability and the availability ratio have been provided.

The macroeconomic variables include, among others: the GDP, the consumer price indices, the interest rates and the unemployment rates. Due to the incompleteness of the individual variables, the number of observations per variable ranges between 3,520 observations (for interest rates) to 220 observations (for a change of GDP forecast by the IMF). For some regions, individual variables are not available, either because the property type is not documented or because the data providers do not cover those specific markets. For instance, the consumer confidence indicator from the OECD is not available for all countries. A combined variable with national-specific and OECD values has been constructed.

Due to friction in both datasets, data modifications were necessary. First, the property variables have been harmonized in terms of measures, frequency and currency towards a monthly square-metre EUR value.[5]

On the macroeconomic side, GDP values have been recorded in different scales and have been harmonized to multiples of millions.

Table 8:12 in the Appendix reports all acronyms and Table 8:13 and Table 8:14 provide the descriptive statistics for the used variables.

3.5.1 GOOGLE TRENDS DATA

The collected data from Google Trends is worth descri*bing* in more detail. The search volume data is available from 2004 onwards. Google Trends allows a detailed look at searches within different regions ranging from an international search down to a regional search. According to the provider, the data is based on the analysis of Google web searches over a specified period of time. However, the provided values are only given as normalized values of all searches for the specific search word within the same location at the same time.

Search words with a low volume and repeated searches from single individuals are excluded. The provided data is adjusted for a better comparison between different terms. These results are scaled to a range from 0 to 100. Nevertheless, the manipulation of the data has been criticized before by scholars, who would prefer actual search volumes and the possibility of accessing the subsequent searches and clicks of individuals to get a clearer picture of their behaviour.

Besides the possibility of analysing different search terms in different regions and at different points in time, the application offers the chance to search within different categories. One of these categories is 'Property' (category ID: 0–29).[6] The categorical filter function eliminates different meanings of words, for better and clearer results. However, Google does not explain how it knows that certain words have been searched within this category since the "normal" Google Search does not offer such a pre-filtered option. Dietzel et al. (2014) explain that the categorization is based on individual search behaviour. Each search is placed into a framework of searches before and after the specific search. According to this, a series of

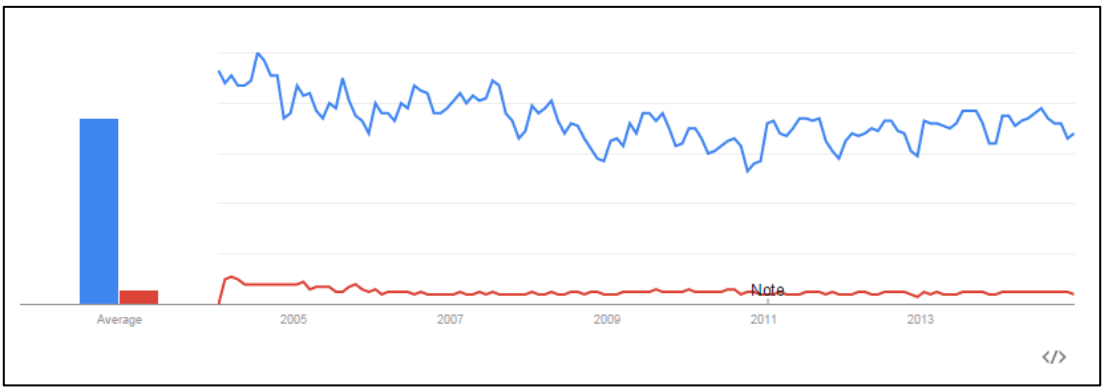
[5] Monetary values recorded in their national currency have been transformed into euros, which was done with the help of historic exchange rates.
[6] The source code of the Google Trends webpage uses those codes for each of the categories.

searches with real estate related search terms would force the underlying algorithm to place searches within the property category.[7] The category comprises further sub-categories: apartments & residential rentals, commercial & investment real property, property development, property inspections & appraisals, property management, real estate agencies, real estate listings, and timeshares & vacation properties.

The dataset for this analysis comprises 80 regions within 24 countries in Europe, including Turkey and the Russian Federation. In comparison to other parts of the world, Europe is characterized by a variety of different languages in a relatively small area. It is advised to perform some simple searches in advance to identify the most optimal way of extracting the data from the online tool. For instance, the word "office" will produce results for the U.K. It can further be used for other countries within Europe and will produce results as well since English is a universal language. However, a German person is more likely to use the German term "Büro". Comparing both searches a difference in the results can be observed.

The following three figures illustrate the search process for the terms "office" and "Büro" and their differences in the provided results.

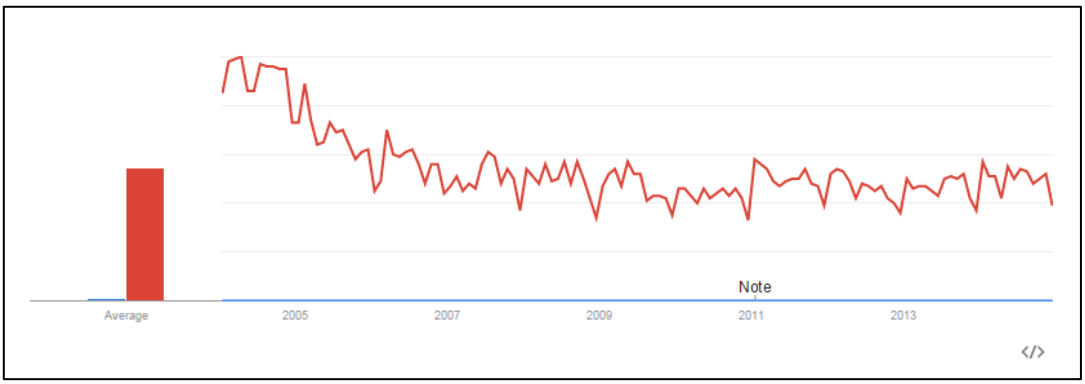
Figure 3:7 - Google Trends - "office"



Note 3.15: Comparison of the term "office" between the U.K. (blue) and Germany (red),8

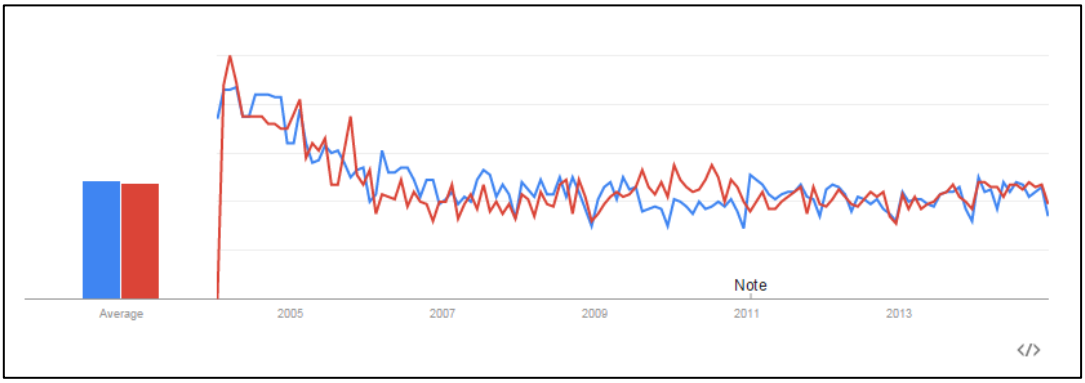
[7] Unfortunately, the authors do not explain where they get this information. Up to this point, I have not been able to get in contact with Google about this and other questions. Google does not offer any service line for GT and emails remain unanswered.
[8] The source for all subsequent graphs/ maps is Google Trends.

Figure 3:8 - Google Trends - "Büro"



Note 3.16: Comparison of the term "Büro" between the U.K. (blue) and Germany (red).

Figure 3:9 - Google Trends - "Büro" vs. "office"



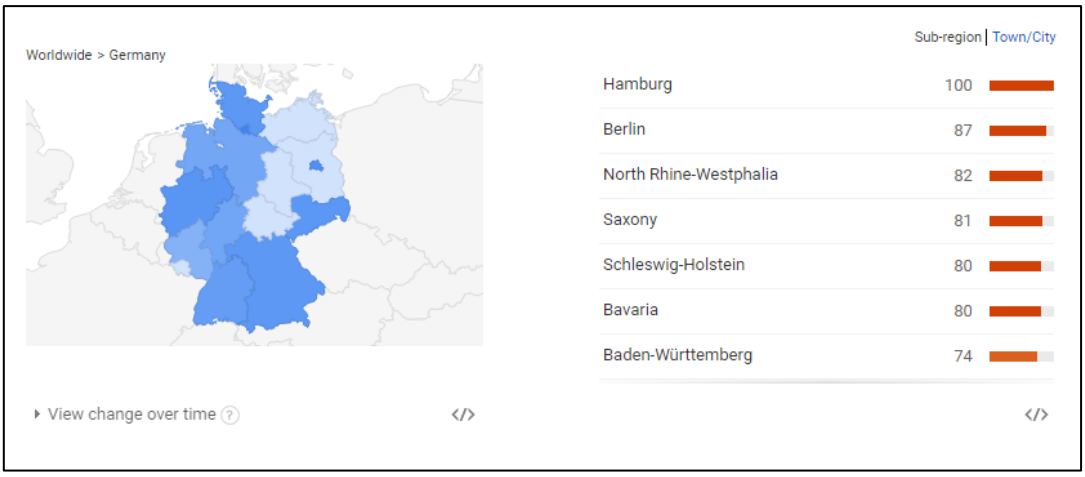
Note 3.17: Comparison between the terms "Büro" (blue) and "office" (red) for Germany

This leads to the fact that the search words need to be translated into the country-specific language. A list of all used words is provided in the Appendix (Table 8:15). Table 8:16 further provides the total score of search words for each city region. For some city regions, only some search words have generated a result.

Besides this language issue, the online tool is limited in the way the data is provided. I assume that location-specific data are more suitable in a real estate context. Therefore the best solution would be to collect the data at a city level. Nevertheless, Google Trends does not offer this option. It is possible to filter for regions within a country, such as the Federal States in Germany; i.e. Berlin, Bavaria, Saxony (Figure 3:10) or the country parts of the United Kingdom (England, Scotland, Northern Ireland and Wales). From there the options are limited. The tool

offers a list of cities with the corresponding share of searches as part of the regional searches (Figure 3:11). However, this share is related to the highest search volume among the cities. Unfortunately, there is no chance of extending the given list to see all cities within the region. Therefore, some cities are not displayed, and a data collection is impossible.

Figure 3:10 - Google Trends - Regional interest



Note 3.18: Regional interest of "Büro' within the Federal States of Germany

Figure 3:11 - Google Trends - City list

Note 3.19: List of cities with the highest search volume for the term "Büro" in relation to each other

Another issue which needs to be addressed is the pure focus on the city and on the region. This might not meet the actual search behaviour. It further excludes the impact of other national

and international investors. Cities such as London, Paris or Frankfurt are probably driven to a significant extent by international investors. National and international interests have been considered within the city-specific data.

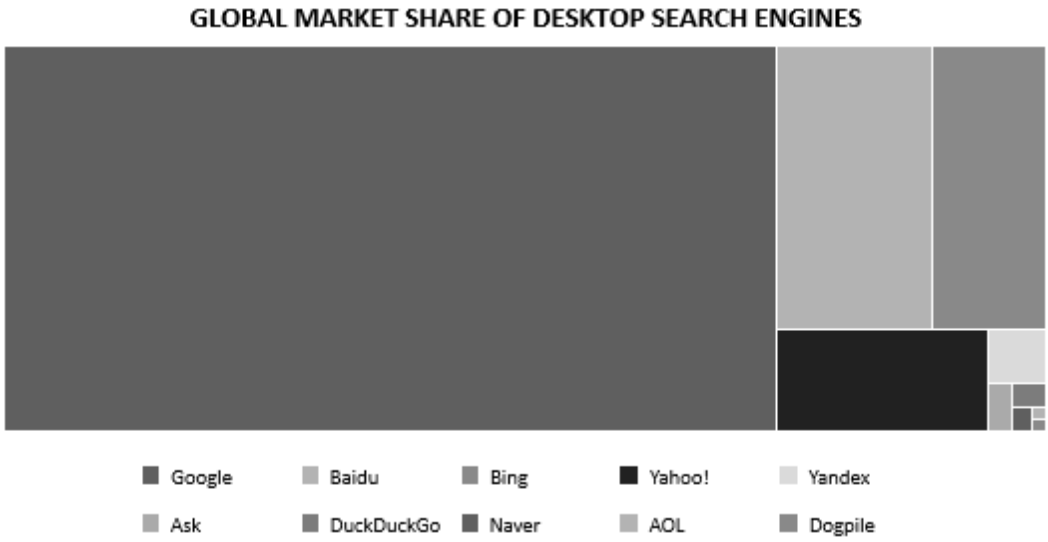
This leads to the question as to how people use the online tool for information mining. Investors or tenants who search for new opportunities or spaces may search first in general, but as soon as they have decided on where they want to go, they are more likely to add a specific city name to their search.

It could be argued that an investor who is interested in buying office space in London will not just Google "*office space*" but "*office space (in) London*". This should return a worldwide map of interest. Nevertheless, the given result for this search in the category "*Property*" in the time between "January 2004 and December 2014" only returns results for England London, based on a worldwide search. One possible explanation would be that the market is not attractive to international or national investors.

Another explanation could be the dense network of real estate service firms. It is unlikely that any investor in person starts to search for an office property on its own. It is more likely that sellers and buyers rely on professionals and their networks. Those professionals are based in those cities, and they may generate these search results.

The assumption that Google might not be used for those specific searches can be denied, based on the given market share of desktop search engines on a global scale (Figure 3:12).

Figure 3:12 - Global market share of desktop search engines[9]



Note 3.20: The figure illustrates the global market share of desktop engines in a worldwide comparison.

To summarize, the online tool offers potential to extract the thoughts of millions of people and the sentiment of the markets. However, the data extractions need to be prepared with care, since a sole focus on regions or cities might not cover the entire picture.

3.5.1.1 CONSTRUCTION OF THE CITY-REGION SPECIFIC GOOGLE TRENDS SERIES

For the construction of the city-specific sentiment measures, I have downloaded the data from the Google Trends website. During that process, I encountered some inconsistencies which I would like to present here.

The displayed graph on the Google Trends page is shown in monthly figures. However, after downloading the file, the results are sometimes shown in weekly figures. It is also possible that both time series do not match. Google does not explain this. For the data collection, a modified version of the R - package GOOGLE TRENDS by Okugami (2013) was used.

Since Google only displays results on a regional level, I have used the list of top cities to calculate a regional indicator. According to Google, the "number represents search volume relative to the highest point on the map which is always 100". These numbers have been used

[9] Source: https://netmarketshare.com, accessed 3 March 2018.

as a percentage share for the specific cities. In cases where the region matched the city, I have used the unchanged score.

The Google Trends results show different intensities for different countries. The number of city regions has been reduced. Some city regions such as Klaipeda, Kaunas, Kyiv, Tallinn and Vilnius have been removed since the available data for those cities was not able to generate any sentiment index. The remaining 75 city regions showed more satisfying and promising results.

The focus on the property category has lowered the possible number of results for the specific search terms. In addition, the results have been limited by the number of search words per search and by the focus on regions. For each region, a set of 90 search terms, which are all related to the commercial real estate market, have been used. Besides more general terms such as “rent” or “office”, the leading service firms and a list of larger European Banks have been included. To cover international interest, a worldwide search with the city name within the property category was performed. The list of words and their frequency can be found in Table 8:15 in the Appendix.

The total amount of search results per city region ranges between 4 (Triangle Area (DK), Malmö (Swe) and Geneva (CH)) and 57 (London (U.K.)). The individual search words scored for each region between 0 and 51 times, though no results were presented for eight search terms (a number of banks and international real estate companies). The Google Trends index for 20 city regions is built out of less than ten search terms.

Besides this, some countries seem not to be covered by the property category at all. For the Czech Republic, Finland, Latvia, Luxembourg, Norway and Romania the general search was used. Meaning that all searches on Google using the search terms have been considered. However, this incorporates noise since not all searches can be directly linked to real estate.

Another reason for the low number of results can be found in translation. Google Translate has been used for all languages.

The following list shows further irregularities in the data collection:

IRELAND

- no region-adjustment possible

SWITZERLAND

- Zurich is placed in the German-speaking part of Switzerland: Canton of Zurich

GERMANY

- Hamburg and Berlin as Federal States seem to be too small to provide sufficient data. Only three out of 46 terms have shown any results. Nevertheless, searching the terms on a national level, Hamburg and Berlin as cities produce more results.

- Berlin & Hamburg: the term “Schulden” (debt) does produce results. However, Google Trends (GT) does not provide any cities where those results have been generated. The results are given as a share of the 16 federal states.

- Bavaria: the term “Darlehen” (mortgage) does produce results for Bavaria. However, there is no share for cities given. The result has been set equal to the overall Bavarian result, based on other results.

CZECH REPUBLIC

- None of the terms has produced any results in the property category. The general search was used. Prague will therefore not fully mirror the real estate sentiment and will incorporate noise.

DENMARK

- Since most of the parts of the triangle area are located in the south of Denmark, the results of the Syddanmark region were used. Those cities which are part of the triangle area (Billund, Fredericia, Vejle, Kolding, Middelfart and Vejen) have been used to generate an average of the region.

- To cover the international interest for the specific property market, I have included a worldwide search for the specific city or region within the property category.

FINLAND

- GT does not offer the property filter function for Finland. To generate data, I have decided to use all categories instead.
- The same applies to the search of the city name in the property context on a worldwide search. It does not seem logical to use the overall search in all categories, because the noise will be too large.

FRANCE

- There is no option to select individual districts of a city, which is a shortcoming of the tool. Furthermore, unreported tests of the worldwide search of the individual districts or areas in the property category have not produced any results.

LATVIA

- GT does not offer the property filter function for Latvia. To generate data, the general search was used.
- Latvia, in comparison to all the other countries within this study, shows the most significant potential in terms of getting fine graded geographical data.

LUXEMBOURG, NORWAY AND ROMANIA

- GT does not offer the property filter function for those countries. To generate data, the general search was used.

3.6 RESULTS

3.6.1 SENTIMENT COMPARISON

As many European countries lack a direct real estate specific sentiment measure, the present study aims to construct close substitutes. The relevance of these indirect measures in models of yields is formally examined within the panel model. Prior to that, it is of interest to get an idea of how closely the alternative indirect measures correspond to direct measures. Given the lack of complete direct measures in Europe, except the U.K., we focus on the London West End market as a case study. In the U.K., RICS (Royal Institution of Chartered Surveyors) have run an established sentiment survey for years.[10] We compare the four indirect indicators (macroeconomic, office, retail and GT) to three RICS sentiment metrics, namely: "Sales and Rental levels" for commercial real estate in London in the next quarter, "Sales and Rental levels" for offices in London in the next quarter, and "Sales and Rental levels" for retail in London in the next quarter. Respondent firms are asked whether sales and rents will over the next quarter: rise, remain similar or fall in relation to the current quarter.

Table 3:10 - Correlation analysis

	U.K. RICS property survey: sales & rental levels-London, next qtr	U.K. RICS survey: office sales & rent levels-London, next qtr nadj	U.K. RICS SURVEY: retail sales & rent levels-London, next qtr nadj
ME sentiment	0.347	0.350	0.279
Google Trends	0.325	0.310	0.269
Office sentiment	0.785	0.766	-
Retail sentiment	0.740	-	0.621

Note 3.21: The table illustrates the correlation between the constructed sentiment measures and the U.K. RICS sentiment surveys.

Table 3:10 shows that the macroeconomic sentiment measure (ME sentiment) has a correlation of 0.347 with the RICS all commercial survey measure. For the office measure, this value increases slightly (0.350) but drops for the retail measure (0.270). This can be seen as a weak correlation. The online search volume measure shows a comparable correlation to the three indicators. The correlation ranges between 0.269 and 0.325.

[10] I have chosen the London West End market, since it provides both the office and the retail market data for the comparison.

On the other hand, the real estate specific indicators exhibit a much stronger correlation of 0.785 and 0.740 for the overall direct sentiment measure. This correlation, unfortunately, drops when it comes to the two more property-type-specific RICS measures. This means that both measures are able to capture some sentiment in the London West End real estate market, an encouraging finding since they nearly perform as well as the direct sentiment proxy.

The macroeconomic and GT measures do not show a high correlation with the RICS surveys, though these correlations are still statistically significant and hence they might pick up some of the sentiment driving real estate markets.

3.6.2 TEST FOR STATIONARITY

Table 3:11 presents the results for the unit root test of all variables used in this analysis. Several tests for stationarity for panel datasets are possible (i.e. the Hadri Lagrange multiplier, the Im-Pesaran-Shin, the Levin-Lin-Chu, the Harris-Tzavalis test). Since the dataset has missing observations for some variables at certain points, the whole dataset can be classified as unbalanced. Therefore, I used Fisher's test for unit roots. The test is designed for unbalanced panel datasets. In general, Fisher's test combines the p-values from N independent unit root tests. Based on the p-values, the test assumes that all series are non-stationary under the null hypothesis. The alternative hypothesis states that at least one series in the panel is stationary. The test allows to specify either the use of the Augmented Dickey Fuller test or the Phillips-Perron unit-root test. The test results suggest, that there is no unit root present and all variables are stationary.

Table 3:11 - Fisher's Unit root test

Label	chi2	Prob > chi2
Office yield	296.8479	0.0000
Retail yield	170.7369	0.0007
Expected_rent_office	187.4232	0.0344
Expected_rent_retail	171.1816	0.0004
Government Bond Rate	416.9408	0.0000
Risk premium	764.4071	0.0000
Macroeconomic sentiment	408.7542	0.0000
Macroeconomic sentiment (Kaiser criterion)	366.6970	0.0000
Macroeconomic sentiment (PCA)	482.7691	0.0000
Office sentiment*	186.2322	0.0000
Retail sentiment*	209.2993	0.0000
Office sentiment (II)	294.2253	0.0000
Property sentiment (I)	824.4145	0.0000
Property sentiment (II)*	656.5825	0.0000
Google Trends	400.0766	0.0000

Note 3.22 - The table presents the individual results of the Fisher unit root test for the different variables used in this analysis. The test has been performed with the consideration of a total number of 4 lags and a drift. For the office, retail and property sentiment (II) I used an older version of the test in STATA (xtfisher). Reasons are that the panels with those sentiment measures did not converge under the xtunitroot option.

3.6.3 EVALUATION OF THE SENTIMENT IMPACT

The results of estimating the yield models with and without indirect sentiment measures for the office sector are given in Table 3:12. In the base model, all variables, except the 10-year government bond rate (5%), are statistically significant at the 1% level and signed as expected. The three proxies for sentiment are introduced individually into the panel model and are statistically significant at the 1% level. The negative sign is in accordance with the expectations. In the sentiment measure, a higher value indicates a stronger sentiment and hence a lower yield. In the model containing the ME sentiment indicator, the rent variable is only significant at a 10% level, and it takes a positive sign, counter-intuitively. All other components remain highly significant and show the expected signs. For the office and the Google Trends model, the government bond rate has a significance of respectively 10% and 5%.

Table 3:12 – Panel regression results: office yield model

Dependent variable office yield				
Variables	Base model	ME sentiment	Office sentiment	ZGT
Expected_rent_office	-0.120***	0.056*	-0.181***	-0.126***
	[0.028]	[0.033]	[0.035]	[0.028]
Government bond	0.020**	0.025***	0.022*	0.020**
	[0.009]	[0.010]	[0.013]	[0.009]
Risk premium	0.024***	0.021***	0.029***	0.025***
	[0.002]	[0.002]	[0.003]	[0.002]
ME sentiment		-0.214***		
		[0.022]		
Office sentiment			-0.102***	
			[0.017]	
Standardized values of (GT)				-0.037***
				[0.009]
Regional fixed effects	Omitted from this output			
Constant	5.803***	5.884***	5.721***	5.818***
	[0.130]	[0.097]	[0.380]	[0.118]
Observations	2,802	2,575	1,496	2,802
Number of cid	69	65	58	69
Correlation coefficient for the actual and fitted value (goodness of fit)	0.867	0.880	0.827	0.871
χ^2	1,896	2,939	2,491	2,288
Df	71	68	61	72

Standard errors in brackets

*** p<0.01, ** p<0.05, * p<0.1

Note 3.23 - The table shows the comparison between the base model and the three different sentiment yield models. The dependent variable is the office yield for the estimation period from 2004q1 to 2014q4. The city fixed effects have been omitted from this report. Amsterdam is the reference region for the output presented above. The omitted regional effects can be found in the Appendix (Table 8:17 to Table 8:20).

The chosen model framework does not allow us to construct a distinct measure of fit, such as an R-squared value. I evaluate the models based on the coefficient of correlation between the observed values of the dependent variable and the fitted values of the dependent variable estimated by each model. There are other methods such as different types of cross-validation or chi-square deviance. However, none of the methods is known to be superior.

On the basis of chosen goodness of fit, models with sentiment make some modest contributions to the explanatory power of the base model, except for the office sentiment

model. The correlation coefficient between actual and fitted values for the base model is 0.867. All but the office specific sentiment induced models outperform the base model. The macroeconomic sentiment model reaches a value of 0.880 and performs best in comparison. The office model reached the lowest correlation with 0.827 and failed to outperform the base model. Finally, the model with the online search volume measure shows the second-best results with 0.871.

The base and the GT model use the same number of city regions (69 regions) and number of observations (2,802 observations). This sample size for the model with the macroeconomic sentiment measure drops a little (65 regions; 2,575 observations) whereas the estimation of the model with the office-specific sentiment measure is based on 58 regions and 1,496 observations. This is caused by data availability of the sentiment proxy (IPD total return for office).

Table 3:13 - Panel regression results: retail yield model

Dependent variable logarithm of retail yield				
Variables	Base model	ME sentiment	Retail sentiment	ZGT
Expected_rent_retail	0.008	0.007	0.018	0.004
	[0.020]	[0.025]	[0.013]	[0.020]
Government bond	0.026***	0.020*	-0.007	0.029***
	[0.010]	[0.010]	[0.010]	[0.010]
Risk premium	0.017***	0.013***	0.009***	0.018***
	[0.002]	[0.002]	[0.002]	[0.002]
ME sentiment		-0.154***		
		[0.021]		
Retail sentiment			-0.808***	
			[0.074]	
Standardized values of (GT)				-0.031***
				[0.009]
Regional fixed effects	Omitted from this output			
Constant	4.408***	4.480***	3.909***	4.397***
	[0.221]	[0.205]	[0.235]	[0.197]
Observations	1,975	1,812	1,629	1,975
Number of cid	51	47	46	51
Correlation coefficient for the actual and fitted value (goodness of fit)	0.869	0.879	0.791	0.872
χ^2	1,021	1,013	882	1,210
Df	53	50	49	54

Standard errors in brackets

*** p<0.01, ** p<0.05, * p<0.1

Note 3.24 - The table shows the comparison between the base model and the three different sentiment yield models. The dependent variable is the retail yield for the estimation period from 2004q1 to 2014q4. The city fixed effects have been omitted from this report. Amsterdam is the reference region for the output presented above. The omitted regional effects can be found in the Appendix Table 8:21 to Table 8:23.

Table 3:13 reports the results for the retail models. Overall the results for the retail side are slightly weaker. It is found that the rent variable for all four models is insignificant. The ten-year government bond rate (risk-free rate) is also insignificant for the retail-specific model. All remaining variables, especially the sentiment measures, are highly significant at the 1% level. The sentiment measures further show the expected negative sign.

Nearly all sentiment induced models outperform the base model (0.869) given the constructed pseudo-measure of fit. The ME sentiment model reaches the highest value with

0.879, followed by the online search volume measure with 0.872. Again, the property specific measure (0.791) fails to provide additional explanatory power to the yield model.

Regarding the number of observations and regions within the different models, we see that only 51 regions are included (47 regions for the ME sentiment model and 46 regions for the property-specific model). Again, this is caused by data availability for the retail market.

SUMMARY

I have found that indirect sentiment indicators constructed in this study are statistically significant variables when included in a base panel model for office or retail yields. The contribution to the base model is marginal to moderate on the basis of the goodness of fit statistic I have used.

The macroeconomic measure has produced the best result for both yield models. This can be seen as a confirmation of the described method of Baker and Wurgler (2007). The property-specific models both failed to outperform the base model and did not provide any additional explanatory power to the standard model. The two property specific indicators are only orthogonalized against one other component. Hence these sector-specific indicators are not filtered sufficiently to extract a pure sentiment component. It can also be argued that the property yield is as suggested in the literature subject to macroeconomic influences and sentiment.

The online search volume indicator has produced the second-best result for both models. This confirms that the easy to use measure provides additional knowledge and should be considered during the modelling process.

3.6.4 FORECAST

The results presented in the previous section are encouraging in the sense that the constructed sentiment proxies have a place and at least should be considered in yield models. I will further assess their validity through an *ex-post* forecast evaluation.

I perform a four-quarter forecast for the period from 2013q1 to 2013q4. Each model is estimated until 2012q4, and both office and retail yield models are forecast for the subsequent four quarters.

Table 3:14 - Forecast evaluation (office models)

	Mean forecast error	Mean absolute error	Mean squared error	Root mean squared error	Theil's U1	Theil's U2	C-statistic
Base Model	-0.501	0.678	1.953	1.397	0.114	1.765	2.118
ME Sentiment	-0.496	0.682	2.305	1.518	0.128	1.900	2.610
Office Sentiment	-0.225	0.358	0.176	0.420	0.039	2.070	3.286
Google Trends	-0.469	0.663	1.948	1.395	0.115	0.505	-0.744

Note 3.25: The table shows the forecast evaluation for the office yield model with the three corresponding sentiment indicators. The columns show the different evaluation measures for the periodic forecast from 2013q1 to 2013q4 on a panel-wide basis.

Table 3:14 illustrates the results of the office yield model. All four models (base and the three sentiment models) show bias in this four-quarter forecasting period as the mean error is not zero. All models have a negative mean forecast error. Therefore, the forecasts tend to be higher than the actual values. Each of the models over-predicts office yield. The office sentiment model has the lowest mean absolute error, mean squared error and root mean square error. The online search volume model ranks second, which means that only the macroeconomic model does not outperform the base model.

Theil's inequality coefficient for all four models is below 0.2 – suggesting good forecast capacity – with the office sentiment model having the lowest value. To check whether the models are able to produce better results than a naïve forecast, I use the yield values of 2012q4 for the next four quarters. The base, the ME sentiment and the office sentiment models have a Theil's U2 value of above one, while only the GT model shows a value below one (Table 3:14). This suggests that the latter model produces better results than a naïve forecast. The same accounts for the last calculated measure, the C-statistic. Only the GT model shows a value below zero, which indicates that the model is able to outperform a naïve forecast on a panel-wide scale.

To conclude, the model with the ME indicator fails to outperform the base model. The office specific measure initially has shown a lower goodness of fit value, yet produced a better result in the forecast evaluation, which could be a period-specific observation.

Table 3:15 - Regional forecast evaluation: office, base model I

Base Model	Ams	Ant	Arh	Bar	Ber	Bir	Bri	Bru	Buc	Bud	Car	Cop	Cor	Dub	Dus	Edi	Fra	Gal
Mean forecast error	-0.087	0.063	-0.432	0.528	-0.380	0.122	0.171	0.047	0.424	0.322	0.166	-0.334	-0.123	0.518	-0.151	-0.016	-0.249	-0.499
Mean absolute error	0.166	0.063	0.432	0.528	0.380	0.232	0.212	0.073	0.424	0.322	0.166	0.334	0.139	0.529	0.151	0.289	0.249	0.499
Mean squared error	0.041	0.005	0.190	0.289	0.145	0.055	0.060	0.009	0.242	0.108	0.038	0.114	0.057	0.390	0.025	0.115	0.065	0.249
Root mean squared error	0.202	0.075	0.436	0.538	0.381	0.236	0.245	0.094	0.492	0.329	0.197	0.338	0.239	0.624	0.159	0.340	0.255	0.499
Theil's U1	0.016	0.005	0.038	0.046	0.037	0.019	0.019	0.007	0.031	0.021	0.015	0.032	0.014	0.051	0.015	0.028	0.024	0.027
Theil's U2	0.598	-	-	3.804	5.079	0.926	0.879	0.546	1.969	-	1.115	-	0.957	0.620	2.248	0.594	5.902	0.999
C-statistic	-0.641	-	-	13.472	24.800	-0.141	-0.226	-0.700	2.878	-	0.243	-	-0.084	-0.615	4.054	-0.647	33.837	-0.000

Base Model	Gen	Gla	Goth	Ham	Hel	Ist	IstAC	IstEC	Kra	Lee	Lie	Lim	LonC	LonD	LonM	LonWe	Lux	Lyo
Mean forecast error	-0.750	-0.047	-0.685	-0.650	-0.622	-0.295	-6.099	-6.098	0.187	0.219	-0.796	-0.570	-0.634	-6.170	-0.698	-0.671	-0.161	-0.245
Mean absolute error	0.750	0.289	0.685	0.650	0.622	0.295	6.099	6.098	0.289	0.219	2.336	0.570	0.634	6.170	0.698	0.671	0.161	0.245
Mean squared error	0.563	0.117	0.472	0.425	0.388	0.088	37.207	37.195	0.122	0.071	10.610	0.325	0.414	38.073	0.499	0.461	0.030	0.060
Root mean squared error	0.750	0.343	0.687	0.652	0.623	0.297	6.099	6.098	0.350	0.267	3.257	0.570	0.643	6.170	0.706	0.679	0.173	0.246
Theil's U1	0.096	0.028	0.064	0.062	0.056	0.020	1.000	1.000	0.023	0.021	0.260	0.030	0.062	1.000	0.067	0.076	0.014	0.020
Theil's U2	3.001	0.599	-	4.542	-	-	-	-	0.777	1.395	0.518	0.570	3.640	-	3.997	3.844	1.390	2.462
C-statistic	8.010	-0.640	-	19.634	-	-	-	-	-0.395	0.945	-0.730	-0.674	12.250	-	14.979	13.780	0.932	5.066

Note 3.26: The table presents the regional specific forecast evaluations for the office market for the base model. Those city regions, with no results for the Theil's U2 and the C-statistic, did not show any variation between the last taken observation in 2012q4 and the four chosen quarters of the forecast. Therefore, the naïve forecast value was equal to the actual values in the four subsequent quarters. Calculating the difference between the actual and the naïve forecast has led to zero. Since both measures use the average of the actual minus the naïve forecast squared as a denominator, the calculation has produced an error.

Table 3:16 - Regional forecast evaluation: office, base model II

Base Model	Mad	Mal	Man	Mar	Mil	Moscow	Mun	New	Not	Osl	P20	PCBD	PCW	PIES	PINS	PIS	PISS	PLBBG
Mean forecast error	0.435	-0.387	-0.035	-0.918	0.176	-1.576	-0.252	0.177	0.263	-0.430	-0.813	-0.813	-0.813	-0.332	-0.386	-0.456	-0.517	-0.438
Mean absolute error	0.435	0.387	0.134	0.918	0.176	1.576	0.252	0.215	0.263	0.430	0.813	0.813	0.813	0.332	0.386	0.456	0.517	0.438
Mean squared error	0.194	0.151	0.024	0.848	0.034	2.487	0.065	0.062	0.075	0.187	0.693	0.693	0.693	0.120	0.158	0.212	0.271	0.196
Root mean squared error	0.440	0.389	0.156	0.920	0.184	1.577	0.255	0.250	0.274	0.433	0.832	0.832	0.832	0.346	0.398	0.461	0.521	0.443
Theil's U1	0.038	0.034	0.013	0.071	0.018	0.080	0.027	0.019	0.020	0.039	0.090	0.090	0.090	0.027	0.032	0.038	0.043	0.040
Theil's U2	3.392	-	0.509	-	3.688	-	10.214	0.895	0.829	1.733	2.220	2.220	2.220	1.961	2.254	-	-	1.772
C-statistic	10.509	-	-0.740	-	12.603	-	103.325	-0.198	-0.312	2.003	3.928	3.928	3.928	2.846	4.084	-	-	2.141

Base Model	PLD	POS	PWC	PWCNBS	PWCNL	PWCSBS	PWCSLD	Pra	Rig	Rom	Rot	She	Sto	THg	Tri	Utr	War	Zur
Mean forecast error	-0.023	-0.070	-0.620	-0.114	-0.642	-0.204	0.188	-0.427	-1.036	0.163	-0.172	0.552	-0.659	-0.058	0.207	-0.072	-0.509	-0.811
Mean absolute error	0.161	0.083	0.620	0.207	0.642	0.204	0.188	0.427	1.036	0.163	0.172	0.552	0.659	0.060	0.212	0.076	0.509	0.811
Mean squared error	0.030	0.014	0.416	0.043	0.444	0.046	0.039	0.184	1.116	0.028	0.031	0.305	0.437	0.008	0.061	0.008	0.264	0.670
Root mean squared error	0.175	0.120	0.645	0.207	0.666	0.215	0.199	0.429	1.056	0.168	0.176	0.552	0.661	0.090	0.247	0.090	0.514	0.818
Theil's U1	0.015	0.009	0.061	0.017	0.063	0.019	0.016	0.033	0.064	0.016	0.013	0.039	0.068	0.007	0.020	0.007	0.039	0.103
Theil's U2	1.399	0.683	1.721	1.661	1.777	0.860	-	-	2.440	1.682	1.442	6.381	4.083	1.474	1.141	0.852	4.117	1.816
C-statistic	0.959	-0.532	1.963	1.759	2.158	-0.260	-	-	4.955	1.829	1.080	39.721	15.676	1.173	0.302	-0.273	15.950	2.299

Note 3.27: The table presents the regional specific forecast evaluations for the office market for the base model. Those city regions, with no results for the Theil's U2 and the C-statistic, did not show any variation between the last taken observation in 2012q4 and the four chosen quarters of the forecast. Therefore, the naïve forecast value was equal to the actual values in the four subsequent quarters. Calculating the difference between the actual and the naïve forecast has led to zero. Since both measures use the average of the actual minus the naïve forecast squared as a denominator, the calculation has produced an error.

Table 3:17 - Regional forecast evaluation: office, ME sentiment model I

ME Sentiment	Ams	Ant	Arh	Bar	Ber	Bir	Bri	Bru	Buc	Bud	Car	Cop	Dus	Edi	Fra	Gen	Gla	Goth
Mean forecast error	-0.070	0.044	-0.389	0.276	-0.273	0.221	0.282	0.029	0.394	0.586	0.255	-0.280	-0.109	0.085	-0.164	-0.730	0.029	-0.631
Mean absolute error	0.163	0.048	0.389	0.276	0.273	0.279	0.282	0.085	0.394	0.586	0.255	0.280	0.109	0.313	0.164	0.730	0.313	0.631
Mean squared error	0.037	0.004	0.155	0.087	0.075	0.092	0.117	0.009	0.208	0.347	0.083	0.082	0.016	0.135	0.030	0.534	0.128	0.399
Root mean squared error	0.194	0.066	0.394	0.296	0.274	0.304	0.342	0.099	0.457	0.589	0.289	0.286	0.127	0.367	0.175	0.731	0.358	0.632
Theil's U1	0.015	0.004	0.034	0.024	0.027	0.025	0.027	0.008	0.028	0.039	0.022	0.027	0.012	0.030	0.017	0.094	0.029	0.059
Theil's U2	0.573	-	-	2.094	3.654	1.192	1.226	0.572	1.828	-	1.636	-	1.798	0.641	4.044	2.924	0.625	-
C-statistic	-0.671	-	-	3.386	12.352	0.422	0.503	-0.672	2.342	-	1.679	-	2.234	-0.587	15.360	7.550	-0.608	-

ME Sentiment	Ham	Hel	Ist	IstAC	IstEC	Kra	Lee	Lie	LonC	LonD	LonM	LonWe	Lux	Lyo	Mad	Mal	Man	Mar
Mean forecast error	-0.575	-0.596	-0.264	-6.693	-6.697	0.213	0.308	-0.744	-0.538	-6.388	-0.616	-0.602	0.063	-0.359	0.181	-0.273	0.057	-0.887
Mean absolute error	0.575	0.596	0.264	6.693	6.697	0.333	0.308	2.334	0.538	6.388	0.616	0.602	0.103	0.359	0.181	0.273	0.125	0.887
Mean squared error	0.335	0.356	0.073	44.799	44.861	0.160	0.129	10.323	0.308	40.817	0.398	0.381	0.011	0.130	0.036	0.075	0.033	0.792
Root mean squared error	0.579	0.597	0.271	6.693	6.697	0.400	0.359	3.213	0.555	6.388	0.631	0.617	0.106	0.360	0.192	0.275	0.181	0.890
Theil's U1	0.056	0.053	0.019	1.000	1.000	0.026	0.028	0.257	0.054	1.000	0.060	0.069	0.009	0.029	0.016	0.024	0.015	0.069
Theil's U2	4.033	-	-	-	-	0.889	1.873	0.511	3.142	-	3.569	3.495	0.849	3.609	1.479	-	0.594	-
C-statistic	15.266	-	-	-	-	-0.209	2.509	-0.738	8.877	-	11.743	11.217	-0.277	12.027	1.189	-	-0.647	-

Note 3.28: The table presents the regional specific forecast evaluations for the office market for the macroeconomic sentiment indicator. Those city regions, with no results for the Theil's U2 and the C-statistic, did not show any variation between the last taken observation in 2012q4 and the four chosen quarters of the forecast. Therefore, the naïve forecast value was equal to the actual values in the four subsequent quarters. Calculating the difference between the actual and the naïve forecast has led to zero. Since both measures use the average of the actual minus the naïve forecast squared as a denominator, the calculation has produced an error.

Table 3:18 - Regional forecast evaluation: office, ME sentiment model II

ME Sentiment	Mil	Moscow	Mun	New	Not	Osl	P20	PCBD	PCW	PIES	PINS	PIS	PISS	PLBBG	PLD	POS	PWC	PWCNBS
Mean forecast error	-0.030	-1.295	-0.182	0.275	0.348	-0.423	-0.754	-0.754	-0.755	-0.283	-0.340	-0.414	-0.455	-0.394	0.038	-0.052	-0.574	-0.072
Mean absolute error	0.055	1.295	0.182	0.275	0.348	0.423	0.754	0.754	0.755	0.283	0.340	0.414	0.455	0.394	0.127	0.110	0.574	0.183
Mean squared error	0.004	1.680	0.034	0.113	0.128	0.182	0.612	0.612	0.612	0.097	0.132	0.176	0.211	0.160	0.030	0.019	0.371	0.034
Root mean squared error	0.065	1.296	0.186	0.337	0.357	0.427	0.782	0.782	0.782	0.311	0.364	0.420	0.460	0.400	0.175	0.138	0.609	0.186
Theil's U1	0.006	0.067	0.019	0.026	0.027	0.039	0.085	0.085	0.085	0.024	0.030	0.035	0.038	0.036	0.015	0.010	0.058	0.015
Theil's U2	1.303	-	7.445	1.207	1.081	1.709	2.086	2.086	2.086	1.762	2.059	-	-	1.600	1.407	0.782	1.624	1.492
C-statistic	0.698	-	54.432	0.456	0.170	1.922	3.354	3.354	3.354	2.104	3.243	-	-	1.562	0.980	-0.387	1.639	1.227

ME Sentiment	PWCNL	PWCSBS	PWCSLD	Rig	Rom	Rot	She	Sto	THg	Tri	Utr	War	Zur
Mean forecast error	-0.595	-0.151	0.228	-0.668	0.012	-0.140	0.664	-0.592	-0.018	0.158	-0.029	-0.372	-0.809
Mean absolute error	0.595	0.151	0.228	0.668	0.046	0.140	0.664	0.592	0.065	0.162	0.052	0.372	0.809
Mean squared error	0.396	0.027	0.056	0.492	0.003	0.022	0.443	0.354	0.006	0.037	0.003	0.149	0.670
Root mean squared error	0.630	0.166	0.237	0.701	0.061	0.150	0.665	0.595	0.081	0.194	0.062	0.386	0.818
Theil's U1	0.060	0.014	0.020	0.044	0.005	0.011	0.047	0.062	0.006	0.015	0.004	0.030	0.103
Theil's U2	1.680	0.664	-	1.621	0.616	1.226	7.688	3.673	1.322	0.898	0.584	3.093	1.817
C-statistic	1.822	-0.559	-	1.627	-0.619	0.505	58.113	12.492	0.748	-0.192	-0.657	8.570	2.301

Note 3.29: The table presents the regional specific forecast evaluations for the office market with the macroeconomic sentiment indicator. Those city regions, with no results for the Theil's U2 and the C-statistic, did not show any variation between the last taken observation in 2012q4 and the four chosen quarters of the forecast. Therefore, the naïve forecast value was equal to the actual values in the four subsequent quarters. Calculating the difference between the actual and the naïve forecast has led to zero. Since both measures use the average of the actual minus the naïve forecast squared as a denominator, the calculation has produced an error.

Table 3:19 - Regional forecast evaluation: office, office sentiment model

Office Sentiment	Ber	Fra	Gen	Ham	Lee	LonC	LonWe	Lyo	Mad	Mal	Mar	Mil	Mun	PLD	PWC	PWCNL	PWCSBS	PWCSLD
Mean forecast error	-0.300	-0.169	-0.417	-0.560	0.237	-0.477	-0.558	-0.162	0.461	-0.175	-0.793	0.219	-0.177	0.139	-0.534	-0.504	-0.066	-
Mean absolute error	0.300	0.169	0.417	0.560	0.237	0.477	0.558	0.162	0.461	0.175	0.793	0.219	0.177	0.139	0.534	0.504	0.112	-
Mean squared error	0.090	0.034	0.176	0.315	0.070	0.233	0.317	0.026	0.220	0.037	0.640	0.057	0.035	0.056	0.310	0.279	0.013	-
Root mean squared error	0.301	0.186	0.419	0.561	0.266	0.483	0.563	0.163	0.469	0.193	0.800	0.238	0.187	0.238	0.556	0.528	0.115	-
Theil's U1	0.030	0.018	0.056	0.054	0.021	0.047	0.064	0.013	0.041	0.017	0.062	0.023	0.020	0.021	0.053	0.050	0.010	-
Theil's U2	4.014	4.302	1.678	3.907	1.385	2.734	3.187	1.636	3.614	-	-	4.776	7.513	1.908	1.485	1.408	0.460	-
C-statistic	15.117	17.511	1.817	14.270	0.920	6.477	9.163	1.677	12.061	-	-	21.810	55.450	2.641	1.205	0.984	-0.788	-

Office Sentiment	Zur
Mean forecast error	-0.621
Mean absolute error	0.621
Mean squared error	0.398
Root mean squared error	0.631
Theil's U1	0.081
Theil's U2	1.400
C-statistic	0.961

Note 3.30: The table presents the regional specific forecast evaluations for the office market with the office specific sentiment indicator. Those city regions, with no results for the Theil's U2 and the C-statistic, did not show any variation between the last taken observation in 2012q4 and the four chosen quarters of the forecast. Therefore, the naïve forecast value was equal to the actual values in the four subsequent quarters. Calculating the difference between the actual and the naïve forecast has led to zero. Since both measures use the average of the actual minus the naïve forecast squared as a denominator, the calculation has produced an error.

Table 3:20 - Regional forecast evaluation: office, Google Trends I

Google Trends	Ams	Ant	Arh	Bar	Ber	Bir	Bri	Bru	Buc	Bud	Car	Cop	Dus	Edi	Fra	Gen	Gla	Goth
Mean forecast error	-0.038	0.131	-0.379	0.609	-0.369	0.145	0.199	0.082	0.388	0.409	0.252	-0.277	-0.132	0.043	-0.219	-0.476	-0.674	0.000
Mean absolute error	0.160	0.131	0.379	0.609	0.369	0.239	0.221	0.089	0.388	0.409	0.252	0.277	0.132	0.284	0.219	0.476	0.674	0.284
Mean squared error	0.032	0.019	0.147	0.381	0.137	0.060	0.069	0.015	0.218	0.169	0.072	0.083	0.020	0.113	0.053	0.228	0.456	0.111
Root mean squared error	0.180	0.139	0.384	0.617	0.370	0.245	0.263	0.125	0.467	0.411	0.269	0.288	0.143	0.336	0.232	0.477	0.675	0.333
Theil's U1	0.014	0.009	0.033	0.053	0.036	0.020	0.021	0.010	0.029	0.027	0.020	0.028	0.013	0.027	0.022	0.025	0.088	0.027
Theil's U2	0.593	0.107	1.535	1.015	0.577	0.354	0.313	0.178	0.177	0.195	0.423	0.231	0.110	0.586	0.156	0.191	0.245	0.804
C-statistic	-0.648	-0.988	1.358	0.030	-0.666	-0.874	-0.901	-0.968	-0.968	-0.961	-0.820	-0.946	-0.987	-0.655	-0.975	-0.963	-0.939	-0.352

Google Trends	Ham	Hel	Ist	IstAC	IstEC	Kra	Lee	Lie	Lon	LonC	LonD	LonH	LonM	LonWe	Lux	Lyo	Mad	Mal
Mean forecast error	-0.597	-0.639	-0.533	-0.235	-6.050	-6.048	0.288	0.252	-0.752	-0.566	-0.599	-6.127	-0.667	-0.646	-0.075	-0.170	0.489	-0.336
Mean absolute error	0.597	0.639	0.533	0.235	6.050	6.048	0.325	0.252	2.363	0.566	0.599	6.127	0.667	0.646	0.075	0.170	0.489	0.336
Mean squared error	0.358	0.410	0.295	0.056	36.604	36.586	0.163	0.086	10.575	0.322	0.369	37.547	0.455	0.428	0.008	0.029	0.243	0.115
Root mean squared error	0.598	0.640	0.543	0.237	6.050	6.048	0.403	0.293	3.251	0.567	0.607	6.127	0.674	0.654	0.091	0.170	0.493	0.339
Theil's U1	0.056	0.061	0.049	0.016	1.000	1.000	0.026	0.023	0.260	0.030	0.058	1.000	0.064	0.073	0.007	0.014	0.043	0.029
Theil's U2	0.478	0.506	0.604	0.278	0.968	0.967	0.293	1.613	0.994	0.214	0.474	0.996	0.526	0.322	0.730	0.782	2.220	0.054
C-statistic	-0.770	-0.743	-0.634	-0.922	-0.062	-0.063	-0.914	1.604	-0.010	-0.954	-0.775	-0.007	-0.722	-0.895	-0.467	-0.388	3.930	-0.997

Note 3.31: The table presents the regional specific forecast evaluations for the office market with the online search volume sentiment indicator (Google Trends). Those city regions, with no results for the Theil's U2 and the C-statistic, did not show any variation between the last taken observation in 2012q4 and the four chosen quarters of the forecast. Therefore, the naïve forecast value was equal to the actual values in the four subsequent quarters. Calculating the difference between the actual and the naïve forecast has led to zero. Since both measures use the average of the actual minus the naïve forecast squared as a denominator, the calculation has produced an error.

Table 3:21 - Regional forecast evaluation: office, Google Trends II

Google Trends	Man	Mar	Mil	Moscow	Mun	New	Not	Osl	P20	PCBD	PCW	PIDF	PIES	PINS	PIS	PISS	PLBBG	PLD
Mean forecast error	-0.000	-0.860	0.212	-1.414	-0.219	0.220	0.276	-0.340	-0.755	-0.755	-0.755	-0.272	-0.321	-0.395	-0.460	-0.385	0.036	-0.082
Mean absolute error	0.116	0.860	0.212	1.414	0.219	0.234	0.276	0.340	0.755	0.755	0.755	0.272	0.321	0.395	0.460	0.385	0.153	0.082
Mean squared error	0.022	0.748	0.048	2.001	0.051	0.078	0.081	0.119	0.599	0.599	0.599	0.082	0.111	0.165	0.220	0.156	0.040	0.015
Root mean squared error	0.149	0.865	0.220	1.414	0.225	0.279	0.285	0.346	0.774	0.774	0.774	0.287	0.334	0.406	0.469	0.396	0.202	0.124
Theil's U1	0.012	0.067	0.021	0.072	0.024	0.021	0.021	0.031	0.084	0.084	0.084	0.023	0.027	0.034	0.039	0.036	0.018	0.009
Theil's U2	0.025	0.150	0.037	0.943	0.041	0.143	0.234	0.125	0.233	0.274	0.333	1.627	0.846	2.708	3.131	0.316	0.169	0.195
C-statistic	-0.999	-0.977	-0.998	-0.110	-0.998	-0.979	-0.945	-0.984	-0.945	-0.924	-0.888	1.648	-0.284	6.334	8.805	-0.899	-0.971	-0.962

Google Trends	POS	PWC	PWCNBS	PWCNL	PWCSBS	PWCSLD	Pra	Rig	Rom	Rot	She	Sto	THg	Tri	Utr	War	Zur
Mean forecast error	-0.567	-0.060	-0.590	-0.144	0.241	-0.374	-0.941	0.223	-0.110	0.580	-0.604	0.018	0.203	0.016	-0.383	-0.796	-0.200
Mean absolute error	0.567	0.202	0.590	0.150	0.241	0.374	0.941	0.223	0.110	0.580	0.604	0.069	0.212	0.047	0.383	0.796	0.200
Mean squared error	0.350	0.043	0.377	0.029	0.066	0.140	0.927	0.053	0.013	0.336	0.369	0.006	0.060	0.002	0.150	0.647	0.001
Root mean squared error	0.592	0.207	0.614	0.170	0.258	0.375	0.963	0.230	0.117	0.580	0.607	0.081	0.244	0.053	0.388	0.804	0.042
Theil's U1	0.056	0.017	0.058	0.015	0.021	0.029	0.059	0.022	0.009	0.041	0.063	0.006	0.020	0.004	0.030	0.101	0.029
Theil's U2	0.192	0.070	0.134	0.037	0.086	0.136	0.691	0.063	0.068	0.857	0.173	0.051	0.134	0.032	0.213	0.181	0.356
C-statistic	-0.962	-0.995	-0.981	-0.998	-0.992	-0.981	-0.521	-0.996	-0.995	-0.264	-0.969	-0.997	-0.981	-0.999	-0.954	-0.967	-0.873

Note 3.32: The table presents the regional specific forecast evaluations for the office market with the online search volume sentiment indicator (Google Trends). Those city regions, with no results for the Theil's U2 and the C-statistic, did not show any variation between the last taken observation in 2012q4 and the four chosen quarters of the forecast. Therefore, the naïve forecast value was equal to the actual values in the four subsequent quarters. Calculating the difference between the actual and the naïve forecast has led to zero. Since both measures use the average of the actual minus the naïve forecast squared as a denominator, the calculation has produced an error.

Looking at the individual results in the forecasts for each region (Table 3:15 to Table 3:21), it can be observed that the results differ from region to region. The ME panel model (Table 3:17 and Table 3:18) performs better than the base model (Table 3:15 and Table 3:16) for most city regions such as Dusseldorf, Frankfurt or Edinburgh.

The Google Trends model (Table 3:20 and Table 3:21) shows similar behaviour. Most of the regions outperform the base model when comparing the mean squared error. Table 3:19 illustrates the results for the office sentiment induced models. It can be observed that nearly all regions perform better than the base model (except Manchester, Madrid, Milano, PLD), which is in line with the overall forecast assessment.

Turning to the retail models, a similar picture is drawn. In Table 3:22 all four models produce a negative mean forecast error, indicating that the models over-predict the yields. The base model has a mean absolute error of 0.538. Only the online search volume sentiment indicator produces a slightly lower value.

Table 3:22 - Forecast evaluation (retail model)

	Mean forecast error	Mean absolute error	Mean squared error	Root mean squared error	Theil's U1	Theil's U2	C-statistic
Base Model	-0.372	0.538	0.795	0.891	0.079	0.964	-0.060
Macroeconomic Sentiment	-0.367	0.547	0.817	0.903	0.081	0.982	-0.034
Retail Sentiment	-0.169	0.590	0.939	0.969	0.096	0.961	-0.074
Google Trends	-0.317	0.505	0.739	0.860	0.077	0.935	-0.125

Note 3.33: The table shows the forecast evaluation for the retail yield model with the three corresponding sentiment indicators. The columns show the different evaluation measures for the periodic forecast from 2013q1 to 2013q4 on a panel-wide basis.

Considering the mean squared error and the root mean squared error criteria the Google Trends indicator model takes a lower value than the base model. Regarding Theil's U1, all models produce values lower than 0.20, which is suggestive of good forecast performance. All models outperform naïve forecast according to Theil's U2 and C-statistics.

Again, none of the indicators is able to outperform the base model consistently. Yet, the online search volume indicator shows a decent performance indicating that it is more suitable to use in a yield model. Compared to the macroeconomic and retail-specific models, the online search volume indicator was able to show a lower mean squared error. Given the fact that all

models did produce higher pseudo-goodness of fit values in the general panel model, the reason for the low performance could be due to periodical circumstances.

Table 3:23 - Regional forecast evaluation: retail, base model

Base model	Ams	Ant	Arh	Bar	Ber	Bir	Bri	Bru	Buc	Bud	Car	Cop	Dus	Edi	Fra	Gen	Gla	Goth
Mean forecast error	-0.431	-0.229	-0.502	-0.023	-0.531	0.036	0.621	-0.141	0.397	-0.201	0.166	-0.077	-0.318	-0.171	-0.405	-0.461	-0.088	-0.525
Mean absolute error	0.431	0.229	0.502	0.023	0.531	0.036	0.621	0.141	0.397	0.201	0.166	0.077	0.318	0.243	0.405	0.461	0.199	0.525
Mean squared error	0.189	0.053	0.254	0.001	0.282	0.002	0.400	0.021	0.159	0.045	0.028	0.008	0.102	0.089	0.165	0.213	0.049	0.277
Root mean squared error	0.434	0.231	0.504	0.027	0.531	0.043	0.632	0.145	0.399	0.212	0.168	0.088	0.319	0.298	0.406	0.462	0.222	0.526
Theil's U1	0.049	0.023	0.048	0.002	0.054	0.004	0.054	0.014	0.024	0.014	0.016	0.009	0.034	0.029	0.043	0.062	0.022	0.050
Theil's U2	5.495	-	-	-	9.507	-	3.576	-	-	-	-	-	5.707	0.843	7.265	1.847	0.591	-
C-statistic	29.198	-	-	-	89.377	-	11.786	-	-	-	-	-	31.565	-0.289	51.776	2.410	-0.651	-

Base model	Ham	Hel	Ist	Kra	Lee	Lie	LonWe	Lux	Lyo	Mad	Mal	Man	Mar	Mil	Moscow	Mun	New	Not
Mean forecast error	-0.525	-0.460	-0.987	-0.201	0.407	0.516	-0.769	-0.035	-0.391	0.116	-0.192	0.323	-1.120	-0.020	-2.100	-0.285	0.265	0.099
Mean absolute error	0.525	0.460	0.987	0.201	0.407	0.516	0.769	0.045	0.391	0.116	0.192	0.323	1.120	0.049	2.100	0.285	0.265	0.099
Mean squared error	0.277	0.212	0.976	0.041	0.166	0.267	0.599	0.006	0.156	0.014	0.038	0.112	1.257	0.005	4.409	0.082	0.082	0.022
Root mean squared error	0.526	0.461	0.988	0.203	0.408	0.517	0.774	0.080	0.395	0.116	0.194	0.335	1.121	0.071	2.100	0.287	0.286	0.148
Theil's U1	0.055	0.042	0.073	0.015	0.037	0.052	0.095	0.007	0.040	0.010	0.017	0.032	0.101	0.007	0.095	0.033	0.028	0.014
Theil's U2	9.408	-	-	-	-	0.099	2.340	0.640	1.581	-	-	2.677	4.484	0.571	-	5.129	1.322	0.682
C-statistic	87.508	-	-	-	-	-0.990	4.478	-0.590	1.500	-	-	6.164	19.106	-0.675	-	25.304	0.748	-0.535

Base model	Osl	P20	Pra	Rig	Rom	Rot	Sto	THg	Tri	Utr	War	Zur
Mean forecast error	-0.453	-4.689	-0.515	-1.046	0.045	-0.427	-0.612	-0.479	0.867	-0.562	-0.725	-1.054
Mean absolute error	0.453	4.689	0.515	1.046	0.045	0.427	0.612	0.479	0.867	0.562	0.725	1.054
Mean squared error	0.206	21.993	0.276	1.095	0.004	0.190	0.377	0.241	0.754	0.320	0.527	1.125
Root mean squared error	0.454	4.690	0.525	1.047	0.060	0.436	0.614	0.491	0.868	0.565	0.726	1.061
Theil's U1	0.041	1.000	0.041	0.070	0.005	0.045	0.064	0.051	0.082	0.058	0.057	0.130
Theil's U2	-	-	2.971	-	0.402	2.946	4.640	3.649	-	8.549	-	0.298
C-statistic	-	-	7.826	-	-0.839	7.682	20.525	12.314	-	72.083	-	-0.911

Note 3.34: The table presents the regional specific forecast evaluations for the retail market for the base model. Those city regions, with no results for the Theil's U2 and the C-statistic, did not show any variation between the last taken observation in 2012q4 and the four chosen quarters of the forecast. Therefore, the naïve forecast value was equal to the actual values in the four subsequent quarters. Calculating the difference between the actual and the naïve forecast has led to zero. Since both measures use the average of the actual minus the naïve forecast squared as a denominator, the calculation has produced an error.

Table 3:24 - Regional forecast evaluation: retail, ME sentiment

ME sentiment	Ams	Ant	Arh	Bar	Ber	Bir	Bri	Bru	Buc	Bud	Car	Cop	Dus	Edi	Fra	Gen	Gla	Goth
Mean forecast error	-0.435	-0.278	-0.464	-0.226	-0.447	0.067	0.613	-0.189	0.357	0.414	0.160	-0.028	-0.142	-0.147	-0.347	-0.487	-0.057	-0.469
Mean absolute error	0.435	0.278	0.464	0.226	0.447	0.067	0.613	0.189	0.357	0.414	0.160	0.042	0.142	0.257	0.347	0.487	0.199	0.469
Mean squared error	0.192	0.078	0.217	0.051	0.201	0.006	0.395	0.037	0.127	0.173	0.027	0.002	0.021	0.089	0.122	0.237	0.053	0.220
Root mean squared error	0.438	0.280	0.466	0.226	0.449	0.075	0.628	0.192	0.357	0.415	0.163	0.050	0.146	0.298	0.349	0.487	0.230	0.469
Theil's U1	0.049	0.027	0.045	0.019	0.046	0.008	0.054	0.019	0.021	0.028	0.016	0.005	0.016	0.029	0.037	0.065	0.023	0.045
Theil's U2	5.541	-	-	-	8.026	-	3.553	-	-	-	-	-	2.617	0.843	6.235	1.949	0.612	-
C-statistic	29.700	-	-	-	63.422	-	11.627	-	-	-	-	-	5.848	-0.290	37.880	2.799	-0.625	-

ME sentiment	Ham	Hel	Ist	Kra	Lee	Lie	LonWe	Lux	Lyo	Mad	Mal	Man	Mar	Mil	Moscow	Mun	New	Not
Mean forecast error	-0.439	-0.475	-1.004	-0.275	0.419	0.506	-0.698	0.065	-0.519	-0.105	-0.176	0.354	-1.165	-0.135	-2.292	-0.274	0.273	0.115
Mean absolute error	0.439	0.475	1.004	0.275	0.419	0.506	0.698	0.116	0.519	0.105	0.176	0.354	1.165	0.135	2.292	0.274	0.273	0.115
Mean squared error	0.194	0.226	1.012	0.077	0.177	0.257	0.495	0.014	0.271	0.011	0.032	0.134	1.360	0.022	5.252	0.076	0.091	0.030
Root mean squared error	0.441	0.475	1.006	0.277	0.421	0.507	0.704	0.117	0.521	0.106	0.177	0.366	1.166	0.148	2.292	0.276	0.302	0.172
Theil's U1	0.047	0.043	0.075	0.021	0.038	0.051	0.087	0.011	0.052	0.009	0.016	0.035	0.104	0.013	0.103	0.032	0.029	0.016
Theil's U2	7.886	-	-	-	-	0.098	2.128	0.938	2.083	-	-	2.925	4.665	1.186	-	4.940	1.395	0.796
C-statistic	61.194	-	-	-	-	-0.990	3.530	-0.121	3.337	-	-	7.553	20.761	0.406	-	23.400	0.945	-0.366

ME sentiment	Osl	P20	Pra	Rig	Rom	Rot	Sto	THg	Tri	Utr	War	Zur
Mean forecast error	-0.478	-4.756	-0.515	-0.885	-0.073	-0.455	-0.590	-0.515	0.819	-0.586	-0.644	-1.050
Mean absolute error	0.478	4.756	0.515	0.885	0.078	0.455	0.590	0.515	0.819	0.586	0.644	1.050
Mean squared error	0.230	22.619	0.280	0.783	0.008	0.213	0.349	0.276	0.672	0.347	0.416	1.110
Root mean squared error	0.480	4.756	0.529	0.885	0.088	0.462	0.591	0.525	0.820	0.589	0.645	1.054
Theil's U1	0.044	1.000	0.041	0.059	0.008	0.047	0.061	0.054	0.077	0.061	0.051	0.129
Theil's U2	-	-	2.991	-	0.584	3.121	4.468	3.903	-	8.901	-	0.296
C-statistic	-	-	7.947	-	-0.659	8.742	18.963	14.235	-	78.230	-	-0.913

Note 3.35: The table presents the regional specific forecast evaluations for the retail market with the macroeconomic sentiment indicator. Those city regions, with no results for the Theil's U2 and the C-statistic, did not show any variation between the last taken observation in 2012q4 and the four chosen quarters of the forecast. Therefore, the naïve forecast value was equal to the actual values in the four subsequent quarters. Calculating the difference between the actual and the naïve forecast has led to zero. Since both measures use the average of the actual minus the naïve forecast squared as a denominator, the calculation has produced an error.

Table 3:25 - Regional forecast evaluation: retail, retail sentiment

Retail sentiment	Ams	Ant	Arh	Bar	Ber	Bir	Bri	Bru	Car	Cop	Dus	Edi	Fra	Gen	Gla	Goth	Ham	Kra
Mean forecast error	-0.514	-0.241	-0.578	-0.006	-0.551	0.532	1.105	-0.145	0.669	-0.130	-0.333	0.319	-0.428	-0.466	0.393	-0.045	-0.537	-0.220
Mean absolute error	0.514	0.241	0.578	0.007	0.551	0.532	1.105	0.145	0.669	0.130	0.333	0.319	0.428	0.466	0.393	0.070	0.537	0.220
Mean squared error	0.267	0.059	0.335	0.000	0.304	0.302	1.223	0.022	0.468	0.018	0.111	0.117	0.184	0.218	0.161	0.007	0.289	0.049
Root mean squared error	0.516	0.243	0.579	0.009	0.551	0.550	1.106	0.148	0.684	0.135	0.333	0.341	0.429	0.467	0.401	0.083	0.537	0.222
Theil's U1	0.057	0.024	0.055	0.001	0.056	0.058	0.099	0.015	0.070	0.014	0.036	0.035	0.045	0.062	0.042	0.008	0.056	0.017
Theil's U2	6.532	-	-	-	9.865	-	6.256	-	-	-	5.965	0.966	7.667	1.866	1.070	-	9.610	-
C-statistic	41.671	-	-	-	96.325	-	38.142	-	-	-	34.578	-0.068	57.780	2.483	0.146	-	91.348	-

Retail sentiment	Lee	Lie	LonWe	Lyo	Mad	Mal	Man	Mar	Mil	Mun	New	Not	Osl	P20	Pra	Rom	Rot	Sto
Mean forecast error	0.915	0.582	-0.320	-0.439	0.137	0.306	0.824	-1.177	0.190	-0.300	0.775	0.627	-0.383	-5.037	-0.460	0.209	-0.499	-0.117
Mean absolute error	0.915	0.582	0.320	0.439	0.137	0.306	0.824	1.177	0.190	0.300	0.775	0.627	0.383	5.037	0.460	0.209	0.499	0.117
Mean squared error	0.856	0.341	0.114	0.195	0.019	0.098	0.690	1.387	0.042	0.090	0.609	0.402	0.149	25.368	0.223	0.045	0.256	0.016
Root mean squared error	0.925	0.584	0.337	0.441	0.137	0.314	0.831	1.178	0.205	0.300	0.781	0.634	0.386	5.037	0.472	0.211	0.506	0.126
Theil's U1	0.087	0.059	0.044	0.044	0.012	0.029	0.083	0.105	0.019	0.034	0.079	0.063	0.035	1.000	0.037	0.020	0.052	0.014
Theil's U2	-	0.112	1.020	1.765	-	-	6.647	4.711	1.644	5.375	3.606	2.928	-	-	2.669	1.409	3.421	0.956
C-statistic	-	-0.987	0.040	2.116	-	-	43.188	21.193	1.702	27.891	12.000	7.576	-	-	6.125	0.986	10.703	-0.086

Retail sentiment	THg	Tri	Utr	Zur
Mean forecast error	-0.547	0.819	-0.639	-1.060
Mean absolute error	0.547	0.819	0.639	1.060
Mean squared error	0.310	0.672	0.412	1.134
Root mean squared error	0.557	0.820	0.642	1.065
Theil's U1	0.057	0.077	0.066	0.130
Theil's U2	4.135	-	9.701	0.299
C-statistic	16.095	-	93.106	-0.911

Note 3.36: The table presents the regional specific forecast evaluations for the retail market with the retail-specific sentiment indicator. Those city regions, with no results for the Theil's U2 and the C-statistic, did not show any variation between the last taken observation in 2012q4 and the four chosen quarters of the forecast. Therefore, the naïve forecast value was equal to the actual values in the four subsequent quarters. Calculating the difference between the actual and the naïve forecast has led to zero. Since both measures use the average of the actual minus the naïve forecast squared as a denominator, the calculation has produced an error.

Table 3:26 - Regional forecast evaluation: retail, Google Trends

Google Trends	Ams	Ant	Arh	Bar	Ber	Bir	Bri	Bru	Buc	Bud	Car	Cop	Dus	Edi	Fra	Gen	Gla	Goth
Mean forecast error	-0.366	-0.157	-0.434	0.042	-0.499	0.066	0.639	-0.095	0.374	-0.097	0.236	0.010	-0.278	-0.121	-0.369	-0.445	-0.034	-0.429
Mean absolute error	0.366	0.157	0.434	0.042	0.499	0.066	0.639	0.095	0.374	0.097	0.236	0.044	0.278	0.243	0.369	0.445	0.181	0.429
Mean squared error	0.137	0.026	0.190	0.002	0.250	0.005	0.423	0.011	0.142	0.010	0.056	0.003	0.078	0.074	0.137	0.199	0.043	0.185
Root mean squared error	0.370	0.160	0.436	0.046	0.500	0.069	0.651	0.103	0.377	0.100	0.236	0.057	0.279	0.273	0.370	0.446	0.206	0.430
Theil's U1	0.042	0.016	0.042	0.004	0.051	0.007	0.056	0.010	0.023	0.007	0.023	0.006	0.030	0.027	0.039	0.060	0.021	0.041
Theil's U2	4.685	-	-	-	8.939	-	3.680	-	-	-	-	-	4.994	0.771	6.619	1.784	0.550	-
C-statistic	20.953	-	-	-	78.898	-	12.545	-	-	-	-	-	23.939	-0.406	42.817	2.181	-0.697	-

Google Trends	Ham	Hel	Ist	Kra	Lee	Lie	LonWe	Lux	Lyo	Mad	Mal	Man	Mar	Mil	Moscow	Mun	New	Not
Mean forecast error	-0.491	-0.366	-0.912	-0.158	0.444	0.564	-0.719	0.013	-0.311	0.141	-0.142	0.364	-1.058	0.033	-1.953	-0.232	0.296	0.125
Mean absolute error	0.491	0.366	0.912	0.158	0.444	0.564	0.719	0.068	0.311	0.141	0.142	0.364	1.058	0.071	1.953	0.232	0.296	0.125
Mean squared error	0.242	0.140	0.833	0.026	0.198	0.320	0.524	0.005	0.101	0.020	0.021	0.141	1.124	0.005	3.817	0.055	0.100	0.028
Root mean squared error	0.492	0.374	0.913	0.162	0.445	0.565	0.724	0.072	0.318	0.141	0.145	0.375	1.060	0.071	1.954	0.234	0.316	0.167
Theil's U1	0.052	0.034	0.068	0.012	0.040	0.057	0.089	0.007	0.032	0.012	0.013	0.036	0.096	0.007	0.089	0.027	0.031	0.016
Theil's U2	8.795	-	-	-	-	0.109	2.189	0.574	1.272	-	-	3.001	4.240	0.570	-	4.189	1.459	0.770
C-statistic	76.358	-	-	-	-	-0.988	3.793	-0.671	0.617	-	-	8.005	16.978	-0.675	-	16.547	1.129	-0.407

Google Trends	Osl	P20	Pra	Rig	Rom	Rot	Sto	THg	Tri	Utr	War	Zur
Mean forecast error	-0.368	-4.629	-0.476	-0.940	0.107	-0.367	-0.554	-0.408	0.872	-0.479	-0.605	-1.061
Mean absolute error	0.368	4.629	0.476	0.940	0.107	0.367	0.554	0.408	0.872	0.479	0.605	1.061
Mean squared error	0.136	21.431	0.244	0.887	0.014	0.143	0.309	0.179	0.762	0.233	0.367	1.138
Root mean squared error	0.369	4.629	0.494	0.942	0.116	0.378	0.556	0.423	0.873	0.483	0.606	1.067
Theil's U1	0.034	1.000	0.039	0.063	0.011	0.039	0.058	0.044	0.082	0.050	0.048	0.130
Theil's U2	-	-	2.794	-	0.776	2.555	4.204	3.139	-	7.297	-	0.299
C-statistic	-	-	6.805	-	-0.398	5.529	16.674	8.851	-	52.248	-	-0.910

Note 3.37: The table presents the regional specific forecast evaluations for the retail market with the online search volume sentiment indicator (Google Trends). Those city regions, with no results for the Theil's U2 and the C-statistic, did not show any variation between the last taken observation in 2012q4 and the four chosen quarters of the forecast. Therefore, the naïve forecast value was equal to the actual values in the four subsequent quarters. Calculating the difference between the actual and the naïve forecast has led to zero. Since both measures use the average of the actual minus the naïve forecast squared as a denominator, the calculation has produced an error.

Looking at the regional forecasts (Table 3:23 to Table 3:26) for the retail model the results are now much more diverse. Comparing the mean squared errors for the different models and regions, it can be seen that the base model is outperformed for most of the various regions.

The Google trends model (Table 3:26) especially shows good performance. The results for the retail model on the other hand (Table 3:25) confirm the initial statement, where the base model produces better results. The ME model on the other hand (Table 3:24) outperforms the base model in most of the cases.

3.6.5 ROBUSTNESS CHECKS

The above results have confirmed my initial hypotheses. First, the standard yield model has benefited from the consideration of sentiment. And second, it seems the constructed sentiment indicators have extracted the sentiment from the sentiment proxies. This was shown by the correlation analysis with the RICS direct sentiment measure. This suggests that the statement in Baker and Wurgler (2007) is correct and all imperfect sentiment proxies carry at least some pure sentiment.

In this section, I will perform two robustness checks to validate my findings. First, I will test the constructed sentiment indicators against the other indicators, which I have mentioned before. Further, I will analyse the above dataset in more detail. The dataset consists of a mixture of various countries with different economic strengths. Therefore, I intend to slice the dataset into two parts, where one part will only incorporate economically strong countries, namely Germany, the U.K. and France (GUF). The remaining countries will also be compiled (rEUR). This should reduce the blurring effect by more stronger countries and provide the strength of the sentiment indicators.

3.6.5.1 SENTIMENT COMPARISON: MACROECONOMIC INDICATOR

The two additional macroeconomic sentiment indicators will be added to the yield model to check if they perform in any way better than the indicator which is based on the suggested method. Reasons for their construction have been presented above.

Table 3:27 presents the results of the office yield model. The three methods only differ slightly from each other. The original method shows significant model parameters and a highly significant sentiment measure. The macroeconomic measure based on the Kaiser Criterion

showed an insignificant rent variable and a sentiment coefficient, which is significant at the 10% level.

The sentiment indicator, which has tried to extract the sentiment by PCA of the sentiment proxies, has produced sufficient model parameters, where all model components are highly significant at the 1% level. Compared to the original measure, it can be seen as an improvement, since the rent variable has now the expected negative sign.

Looking at the values of the pseudo-goodness of fit all models outperform the base model. However, it becomes apparent that the original method (0.880) does produce the best results. The Kaiser Criterion has not helped to improve the model and indicator performance (0.868). While the PCA model has produced the best model parameters, it only ranks second, based on the pseudo-goodness of fit (0.873).

Table 3:27 - Robustness check: ME sentiment comparison, office yield

Dependent variable office yield				
Variables	Base model	ME sentiment	ME sentiment (Kaiser Criterion)	ME sentiment (PCA)
Expected_rent_office	-0.120***	0.056*	0.047	-0.164***
	[0.028]	[0.033]	[0.032]	[0.030]
Government bond	0.020**	0.025***	0.026***	0.052***
	[0.009]	[0.010]	[0.010]	[0.010]
Risk premium	0.024***	0.021***	0.024***	0.022***
	[0.002]	[0.002]	[0.002]	[0.002]
ME sentiment		-0.214***		
		[0.022]		
ME sentiment (Kaiser Criterion)			-0.055*	
			[0.031]	
ME sentiment (PCA)				-0.082***
				[0.006]
Regional fixed effects	Omitted from this output			
Constant	5.803***	5.884***	5.778***	5.706***
	[0.130]	[0.097]	[0.115]	[0.117]
Observations	2,802	2,575	2,572	2,710
Number of cid	69	65	65	65
Correlation coefficient for the actual and fitted value (goodness of fit)	0.867	0.880	0.873	0.868
χ^2	1,896	2,939	2,087	2,056
Df	71	68	68	68

Standard errors in brackets

*** p<0.01, ** p<0.05, * p<0.1

Note 3.38: The table illustrates the regression results for the comparison of the different ME sentiment methods for the office market. The results suggest, that the standard method produces the best results. However, both tested methods still outperform the base model.

For the retail model, the results are presented in Table 3:28. The results are in favour of the original macroeconomic measure. Similar to the office model, the macroeconomic measure based on the Kaiser Criterion shows the lowest result. The coefficient of the sentiment measure remains insignificant. The PCA macroeconomic measure on the other hand has a highly significant coefficient at a 1% level. All three models are able to outperform the base model (0.869). The original macroeconomic measure reaches the highest pseudo-R-square value with 0.879, followed by the Kaiser Criterion (0.875). The PCA measure only ranks third in comparison. This is somehow surprising given the highly significant sentiment coefficient.

Table 3:28 - Robustness check: ME sentiment comparison, retail yield

Dependent variable retail yield				
Variables	Base model	ME sentiment	ME sentiment (Kaiser Criterion)	ME sentiment (PCA)
Expected_rent_office	0.008	0.007	0.016	-0.013
	[0.020]	[0.025]	[0.023]	[0.025]
Government bond	0.026***	0.020*	0.025**	0.051***
	[0.010]	[0.010]	[0.010]	[0.011]
Risk premium	0.017***	0.013***	0.016***	0.017***
	[0.002]	[0.002]	[0.002]	[0.002]
ME sentiment		-0.154***		
		[0.021]		
ME sentiment (Kaiser Criterion)			-0.031	
			[0.030]	
ME sentiment (PCA)				-0.051***
				[0.006]
Regional fixed effects		Omitted from this output		
Constant	4.408***	4.480***	4.373***	4.327***
	[0.221]	[0.205]	[0.205]	[0.223]
Observations	1,975	1,812	1,809	1,884
Number of cid	51	47	47	47
Correlation coefficient for the actual and fitted value (goodness of fit)	0.869	0.879	0.875	0.874
χ^2	1021	1013	928.1	928.2
Df	53	50	50	50

Standard errors in brackets

*** p<0.01, ** p<0.05, * p<0.1

Note 3.39: The table illustrates the regression results for the comparison of the different ME sentiment methods for the retail market. The results suggest, that the standard method produces the best results. However, both tested methods still outperform the base model.

In general, it can be said that the newly constructed sentiment indicators show an inferior result. To conclude, there is no additional benefit from changing the recommended method.

3.6.5.2 SENTIMENT COMPARISON: OFFICE INDICATOR

In this section, the additional office indicator is tested. Table 3:29 shows the result. Both office specific sentiment measures fail to outperform the base model. Surprising, however, is the fact that the simpler model does produce better results than the orthogonalized measure (0.840). Yet, the more straightforward measure has weakened the overall performance of the model, since the risk-free rate variable has become insignificant.

Table 3:29 - Robustness check: office sentiment, office yield

Dependent variable office yield			
Variables	Base model	Office sentiment	Office sentiment (rent)
Expected_rent_office	-0.120***	-0.181***	-0.110***
	[0.028]	[0.035]	[0.027]
Government bond	0.020***	0.022*	-0.015
	[0.009]	[0.013]	[0.011]
Risk premium	0.024***	0.029***	0.020***
	[0.002]	[0.003]	[0.002]
Office sentiment		-0.102***	
		[0.017]	
Office sentiment (rent)			-0.617***
			[0.049]
Regional fixed effects	Omitted from this output		
Constant	5.803***	5.721***	5.502***
	[0.130]	[0.380]	[0.133]
Observations	2,802	1,496	2,439
Number of cids	69	58	64
Correlation coefficient for the actual and fitted value (goodness of fit)	0.867	0.827	0.840
χ^2	1,896	2,491	1937
Df	71	61	67

Standard errors in brackets

*** p<0.01, ** p<0.05, * p<0.1

Note 3.40: The table illustrates the regression results for the comparison of the different office sentiment methods for the office market. The results suggest, that both methods fail to outperform the base model.

This test shows that an orthogonalization measure, which considers more factors, produces more robust results. Therefore, the retail measure would have been significantly improved if we had had more property type-specific factors, which could have been removed from the sentiment proxy.

3.6.5.3 SENTIMENT COMPARISON: PROPERTY SPECIFIC INDICATORS

Two other approaches are taken to capture an all-property sentiment. Following the assumption that the office and retail sentiment within the market only represent shares of a more comprehensive commercial real estate sentiment, I first developed an index based on the average of the two property-specific indicators, and second, applied a PCA to the two property indicators to extract a common trend.

Table 3:30 - Correlation analysis

	U.K. RICS property survey: sales & rental levels-London, next qtr	U.K. RICS survey: office sales & rent levels-London, next qtr nadj	U.K. RICS survey: retail sales & rent levels-London, next qtr nadj
ME sentiment	0.347	0.350	0.279
Google Trends	0.325	0.310	0.269
Property sentiment (average)	0.526	0.579	0.387
Property sentiment (PCA)	0.828	0.802	0.729

Note 3.41: The table illustrates the correlation between the constructed sentiment indicators and the direct sentiment indicators for the U.K. market (U.K. RICS surveys indicators).

For both approaches, a significant increase in the correlation towards the RICS property measures is observed (Table 3:30). The correlation coefficients are higher, as documented above. The overall property sentiment, which used the PCA, yields a strong positive correlation.

Table 3:31, however, illustrates that the high correlation does not automatically mean better performance. Compared to the macroeconomic indicator, both models produce slightly worse results. The average property measure shows an insignificant sentiment coefficient, while the PCA property measure has produced an insignificant rent variable. Further, the pseudo-goodness of fit measure suggests that both models fail to outperform the macroeconomic sentiment measure.

Table 3:31 - Robustness check: property sentiment, office yield

Dependent variable office yield				
Variables	Base model	ME sentiment	Property sentiment (average)	Property sentiment (PCA)
Expected_rent_office	-0.120***	0.056*	-0.120***	-0.045
	[0.028]	[0.033]	[0.028]	[0.044]
Government bond	0.020**	0.025***	0.020**	0.033**
	[0.009]	[0.010]	[0.009]	[0.014]
Risk premium	0.024***	0.021***	0.024***	0.028***
	[0.002]	[0.002]	[0.002]	[0.003]
ME sentiment		-0.214***		
		[0.022]		
Property sentiment (average)			-0.012	
			[0.013]	
Property sentiment (PCA)				-0.188***
				[0.035]
Regional fixed effects	Omitted from this output			
Constant	5.803***	5.884***	5.796***	5.620***
	[0.130]	[0.097]	[0.129]	[0.386]
Observations	2,802	2,575	2,802	948
Number of cid	69	65	69	41
Correlation coefficient for the actual and fitted value (goodness of fit)	0.867	0.880	0.867	0.840
χ^2	1,896	2,939	1,933	3,642
Df	71	68	72	44

Standard errors in brackets

*** p<0.01, ** p<0.05, * p<0.1

Note 3.42: The table illustrates the regression results for the comparison of the different property / office sentiment methods for the office market. The results suggest, that both methods fail to outperform the base model as well as the macroeconomic sentiment induced model.

The retail-specific results (Table 3:32) differ slightly. While the average sentiment indicator remains insignificant, the PCA indicator (0.782) does not outperform the macroeconomic indicator (0.879).

Therefore, the produced result is very explicit, and it seems that the recommended method is superior in comparison to the other two tested versions.

Table 3:32 - Robustness check: property sentiment, retail yield

Dependent variable retail yield				
Variables	Base model	ME sentiment	Property sentiment (average)	Property sentiment (PCA)
Expected_rent_office	0.008	0.007	0.008	0.023
	[0.020]	[0.025]	[0.020]	[0.018]
Government bond	0.026***	0.020*	0.026***	0.026**
	[0.010]	[0.010]	[0.010]	[0.013]
Risk premium	0.017***	0.013***	0.017***	0.015***
	[0.002]	[0.002]	[0.002]	[0.003]
ME sentiment		-0.154***		
		[0.021]		
Property sentiment (average)			-0.003	
			[0.013]	
Property sentiment (PCA)				-0.136***
				[0.031]
Regional fixed effects		Omitted from this output		
Constant	4.408***	4.480***	4.402***	4.448***
	[0.221]	[0.205]	[0.218]	[0.409]
Observations	1,975	1,812	1,975	908
Number of cid	51	47	51	40
Correlation coefficient for the actual and fitted value (goodness of fit)	0.869	0.879	0.869	0.782
χ^2	1021	1013	1042	3196
Df	53	50	54	43

Standard errors in brackets

*** p<0.01, ** p<0.05, * p<0.1

Note 3.43: The table illustrates the regression results for the comparison of the different property sentiment methods for the retail market. The results suggest, that both methods fail to outperform the base model as well as the macroeconomic sentiment induced model.

To conclude, the suggested method by Baker and Wurgler (2007) does produce a more robust sentiment indicator than any of the two methods alone. Further, as has become clear, the number of factors which enter the orthogonalization process plays an important role. The more interlinked these factors are, the more of the observable information can be removed.

3.6.5.4 SLICING

Due to the differences in the nature of the various real estate markets, I assume that the initially performed analysis has incorporated some noise. European real estate markets are diverse in terms of transparency and maturity. Western European real estate markets can be assumed to be more established, which should translate into a more robust market system. Here market information, is more or less immediately considered in the pricing. Less established markets will, therefore, be more strongly exposed to sentiment swings.

The dataset has therefore been sliced to examine whether the results are robust and if the sentiment indicators behave differently. The first category includes Germany, the U.K. and France (GUF). Together the three countries provide nearly half of the observations included in the Cushman and Wakefield dataset. The second part incorporates the remaining countries (rEUR).

First, a new set of sentiment indicators is constructed, using the same methods as presented in chapter 3.4.2.3. These indicators are based on the smaller datasets. All new indicators enter the panel yield models.

Table 3:33 - Robustness checks: slicing (GUF), 0ffice yield model

Dependent variable office yield				
Variables	Base model	ME sentiment	Office sentiment	ZGT
Expected_rent_office	-0.158***	0.039	-0.221***	-0.166***
	[0.034]	[0.036]	[0.042]	[0.034]
Government bond	-0.040**	-0.003	-0.003	-0.040**
	[0.017]	[0.015]	[0.020]	[0.017]
Risk premium	0.021***	0.015***	0.024***	0.023***
	[0.003]	[0.004]	[0.004]	[0.003]
ME sentiment		-0.388***		
		[0.047]		
Office sentiment			-0.141***	
			[0.024]	
Standardized values of (GT)				-0.085***
				[0.019]
Regional fixed effects	Omitted from this output			
Constant	4.898***	4.842***	4.803***	4.840***
	[0.147]	[0.104]	[0.117]	[0.124]
Observations	1,527	1,432	979	1,527
Number of cid	35	35	34	35
Correlation coefficient for the actual and fitted value (goodness of fit)	0.74	0.78	0.79	0.76
χ^2	384.6	880.5	599.6	568.1
Df	37	38	37	38

Standard errors in brackets *** p<0.01, ** p<0.05, * p<0.1

Note 3.44: The table shows the comparison between the base model and the three different sentiment yield models. The dependent variable is the office yield for the estimation period from 2004q1 to 2014q4. The city fixed effects have been omitted from this report. Berlin is the reference region for the output presented above. The omitted regional effects can be found in the Appendix Table 8:24 and Table 8:25.

Starting with the GUF dataset, the results for the office sector have changed compared with the full sample results (Table 3:12). Table 3:33 shows that the government bond rate is insignificant in the ME and Office sentiment models, while the expected rent variable loses its significance in the ME sentiment model as well. Sentiment indicators are highly significant with the expected sign across the board.

Measuring the performance of the individual models, the pseudo-goodness of the fit measure has overall dropped down to around 0.74 (base model). Again, the inclusion of sentiment proxies makes a slight contribution. The highest recorded by office sentiment that

pushes the goodness of fit value up to 0.79 followed by the ME sentiment model (0.78). The GT model still outperforms the base model, but only with a marginal contribution and reaches a pseudo-goodness of fit value of 0.76.

Table 3:34 - Robustness checks: slicing (GUF), retail yield model

Dependent variable retail yield				
Variables	Base model	ME Sentiment	Retail Sentiment	ZGT
Expected_rent_retail	-0.014	0.014	-0.086**	-0.015
	[0.038]	[0.041]	[0.042]	[0.038]
Government bond	-0.003	-0.007	-0.049**	0.005
	[0.021]	[0.020]	[0.022]	[0.020]
Risk premium	0.010**	0.001	0.007	0.010**
	[0.004]	[0.005]	[0.005]	[0.004]
ME sentiment		-0.277***		
		[0.050]		
Retail sentiment			-0.652***	
			[0.086]	
Standardized values of (GT)				-0.066***
				[0.023]
Regional fixed effects	Omitted from this output			
Constant	4.943***	5.014***	4.725***	4.889***
	[0.219]	[0.168]	[0.213]	[0.189]
Observations	748	715	695	748
Correlation coefficient for the actual and fitted value (goodness of fit)	17	17	17	17
Number of cid	0.57	0.60	0.62	0.59
χ^2	57.1	129	132.7	86.38
df	19	20	20	20

Standard errors in brackets *** p<0.01, ** p<0.05, * p<0.1

Note 3.45: The table shows the comparison between the base model and the three different sentiment yield models. The dependent variable is the retail yield for the estimation period from 2004q1 to 2014q4. The city fixed effects have been omitted from this report. Berlin is the reference region for the output presented above. The omitted regional effects can be found in the Appendix Table 8:26.

For the retail sector, most of the model components throughout the four models have become insignificant (Table 3:34). All three sentiment indicators are still highly significant with the expected negative sign. The correlation coefficient between the actual and fitted values has dropped dramatically and lies around 0.57 (base model). The macroeconomic indicator, which

was the best performer in the full sample, now only ranks second (0.60). The retail-specific indicator has the highest value with 0.62 and improves upon its performance in the previous analysis.

Table 3:35 - Robustness checks: slicing (rEUR), office yield model

Dependent variable office yield				
Variables	Base model	ME sentiment	Office sentiment	ZGT
Expected_rent_office	-0.079	-0.130	-0.117*	-0.084
	[0.052]	[0.116]	[0.071]	[0.053]
Government bond	0.035***	0.030**	0.015	0.035***
	[0.011]	[0.012]	[0.016]	[0.011]
Risk premium	0.025***	0.025***	0.036***	0.026***
	[0.003]	[0.003]	[0.004]	[0.003]
ME sentiment		0.028		
		[0.023]		
Office sentiment			-0.097***	
			[0.022]	
Standardized values of (GT)				-0.022**
				[0.010]
Regional fixed effects	Omitted from this output			
Constant	5.742***	5.755***	5.647***	5.754***
	[0.137]	[0.136]	[0.354]	[0.130]
Observations	1,275	1,146	517	1,275
Number of cid	34	30	24	34
Correlation coefficient for the actual and fitted value (goodness of fit)	0.903	0.913	0.878	0.904
χ^2	1,366.00	1,228.00	2,161.00	1,495.00
df	36	33	27	37

Standard errors in brackets *** p<0.01, ** p<0.05, * p<0.1

Note 3.46: The table shows the comparison between the base model and the three different sentiment yield models. The dependent variable is the office yield for the estimation period from 2004q1 to 2014q4. The city fixed effects have been omitted from this report. Amsterdam is the reference region for the output presented above. The omitted regional effects can be found in the Appendix Table 8:27 and Table 8:28.

Using the remaining regions as a comparable (rEUR), I have found that the rent variable has become insignificant for all but the office specific sentiment model (Table 3:35). The risk premium is significant at the 1% level. The macroeconomic sentiment indicator is insignificant,

which is surprising. The remaining two indicators are significant at the 1% and 5% level (online search volume).

Regarding the measure of fit, the base model has a correlation coefficient of 0.90. This value has improved in comparison to the full sample. The model containing the office sentiment indicator fails to outperform (correlation coefficient of 0.878) the base model. The GT model shows a goodness of fit score above the result of the base model (0.90). For the macroeconomic model, the indicator is insignificant, which shows the best result with 0.91.

Table 3:36 - Robustness checks: slicing (rEUR), retail yield model

Dependent variable retail yield				
Variables	Base model	ME sentiment	Retail Sentiment	ZGT
Expected_rent_retail	0.013	-0.013	0.029**	0.008
	[0.026]	[0.034]	[0.013]	[0.026]
Government bond	0.031***	0.033***	-0.012	0.035***
	[0.011]	[0.012]	[0.011]	[0.011]
Risk premium	0.020***	0.020***	0.008***	0.021***
	[0.003]	[0.003]	[0.002]	[0.003]
ME sentiment		0.036		
		[0.023]		
Retail Sentiment			-0.822***	
			[0.084]	
Standardized values of (GT)				-0.026***
				[0.010]
Regional fixed effects		Omitted from this output		
Constant	4.359***	4.308***	4.036***	4.344***
	[0.217]	[0.202]	[0.243]	[0.194]
Observations	1,227	1,100	934	1,227
Number of cid	34	30	29	34
Correlation coefficient for the actual and fitted value (goodness of fit)	0.879	0.894	0.832	0.882
χ^2	963.30	894.80	752.80	1,139.00
Df	36	33	32	37

Standard errors in brackets *** p<0.01, ** p<0.05, * p<0.1

Note 3.47: The table shows the comparison between the base model and the three different sentiment yield models. The dependent variable is the retail yield for the estimation period from 2004q1 to 2014q4. The city fixed effects have been omitted from this report. Amsterdam is the reference region for the output presented above. The omitted regional effects can be found in the Appendix Table 8:29 and Table 8:30.

Table 3:36 shows the results for the last group: retail in the non-core countries. The results reveal that the expected rent component has become insignificant for all but the retail-specific model. The risk-free rate and the risk premium are highly significant, while the risk-free rate remains insignificant for the retail-specific model. The macroeconomic indicator is once again insignificant, however, and produces the highest pseudo-measure of fit value (0.89). The other two models carry highly significant sentiment measures, yet only the online search volume measure (0.88) is capable of outperforming the base model (0.88) marginally.

To summarize, it has become apparent that the division of the dataset has changed the behaviour of the constructed sentiment indicators. While in the complete sample the results have been in favour of the macroeconomic indicator, the separation has shown a distinct pattern. Countries in the Western European Union are characterized by more established and more efficient real estate markets leading to more transparent markets with significant information about prices and market developments. Market participants have access and utilize a range of market information. Macroeconomic information still plays a vital role, yet macroeconomic sentiment is processed, and there is no need for a constructed indirect measure.

The office and retail centres in the remaining countries (rEUR) are subject to indirect macroeconomic sentiment. Unfortunately, in both models, the indicator has become insignificant, but macroeconomic sentiment has produced the highest correlation coefficient, clearly demonstrating gaps in incorporating macroeconomic developments within the pricing of properties.

A caveat is necessary here. The second dataset still includes other Western European countries such as the Netherlands, Sweden, Italy and Spain and results reflect the situation in these countries as well, though some signs are obtained as to the sources of sentiment in less developed real estate markets.

The GT indicator especially has proven its usability for the last analysis (rEUR). Compared to the complexity of the methodology of the construction, the GT data is a good substitute, which should be considered within a yield model.

3.7 CONCLUSION

This first analysis has shown that the European real estate market is subject to sentiment. Market participants such as lenders or investors might not always follow a rational path, especially in a market environment where information is scarce. This irrationality can be observed in the relationship of net income from real estate assets (known as NOI – net operating income) and the market price that defines property yields. Market prices and yields may not solely reflect fundamentals in the market as they are also driven by sentiment.

Yield modelling and the role of sentiment that can induce irrationality in property pricing is of interest to various market players. This chapter has outlined the fundamental properties and premises of standard models that existing studies have developed to explain yield adjustments and swings in property values. Scholars stress the importance of the rent growth component in these models since they carry both the regional fixed effects (and hence market idiosyncrasies) as well as the income expectations of market participants. In addition, the widespread view is that shifts in property yields are caused by shifts in underlying market sentiment. Except for the study of Ling et al. (2014), who applied a set of different sentiment measures to the yield model, the field is under-researched.

I have shown that the European real estate market is subject to sentiment. The use of indirect sentiment proxies is a sufficient substitute in the absence of direct sentiment measures. In this way, the contribution to the existing literature is threefold.

The first contribution relates to the sentiment measures. Unlike the measures found in Ling et al. (2014), the focus was set on other sentiment proxies. This was motivated by (i) the underlying idea of Baker and Wurgler (2006) that each imperfect sentiment proxy carries, at least to a certain extent, some pure sentiment and (ii) by data availability.

Forecast evaluations reveal that models incorporating more specific sentiment measures outperform the base model. The property-specific measure produces better results for the office model. The online search volume measure is the only measure which consistently outperformed the base model in the forecast evaluation (panel wide comparison).

Second, the study extends the research area of sentiment-induced yield modelling to the European commercial real estate market. A number of studies focus on the US market, partially triggered by data availability. However, the interest of investors and banks in sudden movements in yields and pricing and the role of market sentiment has grown in Europe following the global financial crisis.

Finally, the more detailed analysis of the dataset has shown that the stage of the real estate market plays a vital role in its sensitivity towards sentiment. While major markets such as Germany, the U.K. or France are less exposed to macroeconomic sentiment swings, the remaining dataset has shown higher goodness of fit measure for this sentiment indicator. This could mean that macroeconomic factors only play a minor role for more established markets since information transparency allows market participants to reflect changes in the economy more or less immediately. This finding is comparable to Mian and Sankaraguruswamy (2012) or Lee et al. (1990) who have analysed the behaviour of young stocks and closed-end funds. According to the authors, small, young, highly volatile and non-dividend paying stocks/funds are more exposed to sentiment shifts.

Besides these satisfying results, a range of questions and obstacles have occurred over the process. First, the usage of sentiment proxies should be treated with caution. Each of the proxies does not measure the sentiment in the first place. It could be seen as controversial whether the presented proxies are able to capture the underlying sentiment. Further, the process of orthogonalization may seem suitable for the extraction of sentiment. However, two questions remain unanswered. First, has the right number of macroeconomic elements been removed from the proxies or is any obvious factor missing? And second, it needs to be questioned whether the process of principal component analysis in its applied form is correct or not. Scholars are discordant with regard to the number of components which should be used. The applied process ignores the Kaiser Criterion, which recommends at least the usage of all components with an eigenvalue above one.

Unfortunately, many European countries do not have a direct sentiment measure. And even if such a measure is present, they are based on different sets of questions. A comparison of the different markets on an international scale with a direct measure is therefore nearly impossible.

The last criticism which needs to be brought forward is on a more theoretical level. The underlying assumption that direct and indirect sentiment indicators are able to measure the sentiment needs to be questioned. Here or in other studies, used sentiment proxies measure economic factors in the first place and do not measure the sentiment within the market. Even though statistical methods such as orthogonalization are able to extract the sentiment, it remains difficult to say whether all economic factors have been removed. However, the advantage of indirect sentiment measures is their universal application. For the direct sentiment measures, the critic goes a step further. Academia assumes that surveys or interviews are able to measure the sentiment in a better way. This seems logical since direct interaction with people

reveals more. However, the construction of such an indicator requires some commitment. The surveys need to be performed on a regular base; and before a well-educated description of the sentiment can be made, a series of these interviews need to be performed. Yet, these surveys consist of questions regarding the expectation of market participants about future developments. Two things are disputable. First, and here I draw the line back to the literature review in Chapter 2, people who read the results of the survey may assume that the results represent the reality and accept them, that they might lead to a change in behaviour. The survey can, therefore, become a self-fulfilling prophecy. The second fact makes me wonder if a survey represents simply a summary of all interviewees. That means that a majority of people did have a certain belief at the moment the survey was conducted. So, the survey cannot be seen as the best source of sentiment, because the sentiment already existed at this point. Therefore, the sentiment is formed at an earlier stage. Therefore, other warnings signs might have been present before and have just led to the formed sentiment. On the other hand, can the interview be seen as an aggregated opinion of a view market participants, while others (the readers) just follow.

So, the question is: what determines the sentiment? Besides personal socialization and other biases, three fields can been identified: a professional framework experience, the interaction with co-workers and the process of information gathering. The first two are difficult to observe, whereas for the last one different source can be used.

The following chapters will pick up this idea and will illustrate how different textual sources, such as market reports and news articles, as information sources, can be used for sentiment analysis.

4 NATURAL LANGUAGE PROCESSING[11]

4.1 INTRODUCTION

The last chapter revealed that sentiment plays a vital role in the real estate market and that it can be measured, even on a European scale. Nevertheless, some shortcomings regarding the direct, the indirect and the hybrid sentiment measures have been identified.

The first set is characterized by a continuous and probably cost-intensive way of construction. This requires a series of repetitions and a significant amount of interview participants. Furthermore, the method may lack the ability to make comparisons between countries, if the structure of the surveys differs. It also needs to be asked whether the expressed sentiment in a survey is the cause or the result of market swings. As stated earlier, it is my belief that the person who is answering the survey questions or is being interviewed has already formed his or her opinion on the market. So, the survey just summarizes the market situation and does not cause the sentiment in the first place. The survey is likely to have a multiplier effect on the broader market and other market participants. So, the overall situation at the moment of the survey must already have been expressed.

The second method, the quantification of sentiment through indirect sentiment measures, reaches its boundaries in two aspects. First the selection of sentiment proxies is rather difficult, and second, the process of orthogonalization leaves open the question as to whether all macroeconomic elements have been removed. Finally, it can be asked whether the residuals of these orthogonalization processes are actually equal to the sentiment of the market. The third and relatively new method of using online search volume data has its disadvantages in the data themselves. Major data providers, such as Google, modify the data before researchers or market participants get access to it. Further, it needs to be asked how we use online search engines. People may gather information about "hot topics" which does not lead to any actual activity within the market.

Given that, the question remains as to how sentiment is formed. Based on personal experience and common sense, three ways of how an opinion can be developed in a

[11] The main parts of this chapter have been transformed into a paper published in the *Journal of Property Investment & Finance*, March 2018, entitled "Measuring Sentiment in Real Estate: A Comparison Study" by S. Heinig and A. Nanda.

professional framework have been identified: experience, interaction with co-workers and information gathering.

Where the first two factors are difficult to observe, without excessive qualitative research, the last one is of interest for the remainder of this thesis. One source of sentiment formation could be the information stored in texts. That information might be the pure sentiment or a pre-stage of it.

Knowledge and experience are used to process the information. These documents are likely to be market reports from service agencies or news articles. All text documents share the advantage of free and easy access. Assuming that market participants want to stay informed, we assume that at least one of the above-mentioned sources is consumed on a regular basis.

This chapter is intended to introduce the field of natural language processing and textual analysis. I analyse a corpus of U.K. market reports with different lexical methods where the documents are sorted based on positively and negatively labelled wordlists. Four different methods are compared. These methods have been used in other fields before. The methods are *AFINN*, *NRC*, *BING* and Topic Modelling.

I like to point out to the reader, that the focus of this chapter is set on the introduction of the methodology and how text documents can be quantified. The modelling part of this study is just used to underline my general assumption. Therefore, I will analyse these new indicators with the help of a simple autoregressive model. My results suggest that quantified market reports incorporate useful information, which can be used to improve total return models. Some of the sentiment indicators produce satisfying results and improve the base model significantly. However, I have identified that an agglomerated analysis of the U.K. market based on the corpus produces better results than a focus on a single market or property types such as London or offices.

A specific London CRE market analysis reaches its limitation due to the low number of documents in the corpus. Regarding the comparison of the four methods, two of them have produced acceptable results throughout this analysis. I am able to conclude that the consideration of the human element expressed in text helps to provide a deeper understanding of market development.

The remainder of this chapter is organized as follows. Another literature review is presented which introduces the field of natural language processing and textual analysis. This is to show

where this method has reached the real estate market. Then, both the theory and the methodology are explained. Finally, the results, as well as a conclusion, are presented.

4.2 LITERATURE REVIEW: TEXTUAL SENTIMENT ANALYSIS

Chapter 2 placed the research topic against the broader background of behavioural finance. I have established in Chapter 3 that real estate is exposed to sentiment and that it is worth applying behavioural finance methods for a deeper understanding of market mechanics. Behavioural finance has put the individual at the centre of interest and has opened the door for other disciplines to interact with the field.

Over recent decades, researchers have identified that sentiment is a suitable way to extract expectations and opinions of individuals. This sentiment is either based on a direct determination via surveys or through the use of suitable macro- or microeconomic proxies. The research in both fields is quite vast, and advantages and disadvantages have been identified.

Another and not yet discussed method will be at the centre of this chapter. Natural language processing (NLP) has been used in a variety of different fields in recent years. NLP enables the researcher to extract the underlying information in a language and in written information in a new way. Due to the rise of the internet and the availability of computers, the volume of information has tremendously increased. NLP offers a unique way to extract sentiment from a corpus of documents.

This section starts with a definition of the field; this should help us to understand what the initial ideas were and how other disciplines have started to use those achievements. The financial market has been using different methods of NLP for some years with success. Due to the popularity of the field, research has increased, and a vast number of studies are available. I hope to give a good overview of those techniques which have been identified as superior. In the interest of the thesis, I will focus in particular on polarity classification and topic modelling. The section will also summarize the almost non-existent research in the real estate market.

4.2.1 NATURAL LANGUAGE PROCESSING: BACKGROUND

Similar to the field of sentiment analysis, research has struggled to define the field of NLP satisfactorily. One main reason might be that NLP has entered many other fields for many different reasons. Therefore, it has become difficult to describe what people actually understand by it.

Montoyo et al. (2012) have identified nine objectives as to why NLP is performed. Categorizing them into four general classes, the authors mention (1) creation of resources for subjectivity analysis, (2) classification of text according to polarity, (3) opinion extraction, and (4) application of sentiment analysis.

Those classes, however, are not clearly separated and offer space for violations. For instance, class 1 is in many cases the starting point of any study in NLP where written documents are gathered and lexicons are developed. In a second step, those resources are used in the classification process (class 2).

Liddy (2001) tries to provide a definition of the field:

"Natural Language Processing is a theoretically motivated range of computational techniques for analysing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications."

Liddy, E. (2001) "Natural Language Processing"; page 2

She further illustrates that the first part of the definition was kept quite vague since NLP can be performed in multiple computational ways. Also, she points out that while naturally occurring texts are either written or oral, they are based on the interaction between humans. Humans are able to process language on multiple levels at once, whereas NLP tools may not be able to present a full picture without difficulties. Referring to human-like language, she included a reference to the origins of the field, the interaction of humans and machines. The author concludes with a vague picture of possible applications. Similar to Montoyo (2012), Liddy (2001) identifies four distinct motivations for the performance of NLP: (1) paraphrase an input text, (2) translation, (3) answer questions about the content of the text and (4) draw inferences from the text.

Topic modelling and sentiment analysis belong to 3 and 4. Similarly, Chowdhury (2003) includes several subcategories which are linked to NLP. Among others, machine translation, natural language text processing and summarization are essential. Pang and Lee (2008) note that the field is vast in its applications and terminology. The authors see this as a standard issue when new fields emerge. Phrases like sentiment analysis and text or opinion mining are heavily used interchangeably. It is not possible to separate the fields from each other. In the opinion of Pang and Lee (2008) both fields, however, are subcategories of subjectivity analysis.

As Liddy (2001) has stated, the field may be new to a variety of other disciplines. However, research started at the end of World War II, when machines were developed to solve more complex tasks. NLP has its origins without a doubt in the field of computer science. Alan Turing (1950) was one of the first to start work in the field. He developed the Turing Test, which tried to find out whether machines are able to display intelligent behaviour. One of the primary needs for this is a working natural language processing system. So, the initial idea of NLP was to improve the communication between humans and machines, and for some years the focus was set on machine translation.

One of the pioneers of machine translation was Chomsky (1957). His work on reproducible grammar helped the field to emerge. Chomsky introduced a variety of notations for splitting up textual content, which have partly remained until today. As has been reviewed by Lees (1957), Chomsky's work as a linguist has helped to build a bridge between linguistics, psychology and computer science. His reproducible grammar method did not aim to define right or wrong but to produce acceptable structures for further interaction. This was based on mathematical algorithms. His initial motivation was set by the fact that humans are unable to know all possible words and sentences, but that we know the structures of the language. This enables us to form hitherto unknown sentences to communicate.

Following this, other researchers were motivated to enter the field, such as Katz and Fodor (1964). In their opinion, grammar only plays a minor role in the understanding of language. They developed a theoretical framework of what language semantics should look like and what parts are needed.

In later years, Chomsky's theory was increasingly criticized. Among other things, the foremost criticism was based on the fact that Chomsky used grammar as a part of the language and did not offer any mechanics for representing or extracting content. Chomsky (1965) presented a better model of transformational grammar.

Following this, we can see that NLP in its early years dealt mainly with the reproduction of language and the analysis of structures. Researchers wanted to link these theories to machines either to develop artificial intelligence or to enable machines to translate text from one language into another.

Without focusing too much on the theoretical background, and keeping in mind that the primary focus of this thesis will lie on sentiment and some more text-based analysis, I conclude that research has developed working NLP systems where human and machines interact on a satisfying level, especially during recent years with the increasing use of the internet and computers. At the same time, the amount of digital content (mostly in the form of texts) has increased massively.

4.2.2 SENTIMENT ANALYSIS

Going one step back to sentiment analysis, I have shown that equity markets are interested in the thoughts and opinions of retail traders. NLP and opinion mining, in particular, can be applied to a variety of fields. Maks and Vossen (2012) state that, among others, product movie and hotel reviews are common fields where these techniques are applied. In many cases, the underlying sentiment or opinions are extracted from blogs or news articles.

The underlying assumption is that individuals are influenced by news and information which surround them. People are able to realize whether texts are positive or negative. According to O'Hare et al. (2009), an increasing number of objects in one document makes it harder for individuals to identify the underlying sentiment. Therefore, manual polarity sorting of documents can't be realized limitlessly. One reason, that people change their behaviour, is the attitude expressed by the author, which influences the reader. This could lead to herding behaviour when people adopt certain opinions out of the fear of being isolated. Maks and Vossen (2012) have identified that subjectivity in texts is a main factor of influence. However, software applications are not able to extract it fully. Biases and attitudes are included in the labelling process of the reader while categorizing the texts as positive or negative. According to the authors, the key for a more profound extraction of sentiment is a finely graded lexicon, in which words are categorized as positive, neutral or negative. This process is known as polarity classification within the literature [O'Hare et al. (2009)] and represents one of the leading applications of NLP.

Picking up on the importance of lexica, Steinberger et al. (2012) have developed an automated translation method for multiple languages. They have criticized the fact that many of the available gold standard word lexicons are only available in English, other languages such as German, Russian or French having been excluded at the time. The proposed technique is triangulation, which is based on two lexica in two different languages which are individually used to translate into the third language. The reason for this method is the vast amount of possible translations of individual words.

Fawcett and Provost (1999) looked at the early stock market applications of text mining and sentiment analysis based on news articles. They showed that news articles and stock prices are linked and that it is possible to establish a warning system for upcoming changes. The authors made clear that their activity monitoring is based on knowledge of machine learning, statistical analysis and database handling. However, a detailed explanation of their procedure is missing.

One further application of NLP for the capital market can be found in Lavrenko et al. (2000). The authors introduced a system which analyses and recommends news articles to the reader based on the idea that those articles will affect the market. The developed system verifies existing financial time series and news articles over the correlation of the content using piecewise linear regression. This approach differs from classical methods where applications and analysts tried to figure out which articles match user interests. Their work is linked to activity monitoring [Fawcett et al. (1999)] and information filtering. This includes the observation of data streams, here in the form of news articles, and generating alarms either positive or negative, which allow users to act appropriately. Their analysis is mainly based on some self-defined word lexica which are developed out of the underlying news articles during the training period.

Other researchers developed similar systems with comparable features. Godbole et al. (2007), for instance, do not entirely focus on the financial news since they consider it either as positive or as negative, but never neutral. This can be traced back to the fact that authors of news articles often end up with one side of the argument. In addition to standard news articles, they extend their analysis to blogs, where the above-mentioned assumption is even stronger than in news articles. In addition, the increasing number of topic-specific blogs underlines their motivation. In contrast to other researchers, they included subjective sentences in their algorithm since news readers are not going to ignore subjective information given by the author. Their study was able to confirm the results of Pang et al. (2002), who used sentiment analysis for movie reviews.

Building on the achievements of Lavrenko et al. (2000) and Fawcett et al. (1999), Fung et al. (2002) recommend a different approach for the weighting of articles which are used for the prediction of market movements. In contrast to Lavrenko et al. (2000), the authors recommend that there should be no exclusion of articles in the training process of the algorithm since this would be contradictory to the Efficient Market Hypothesis (EMH). Following this theory, the market incorporates all available information into the prices immediately. They used a piecewise segmentation algorithm, which discovers trends in the time series and groups the articles into the categories “rise” or “drop”.

In a later study, Fung et al. (2005) also classified articles. However, their approach differed from the studies presented so far and was based on the work of Gidofalvi (2001). They used a training corpus of news articles and aligned certain articles to stock movements. Instead of focusing initially on the articles, they only aligned them when the stock price changed after the publication of an article, an increasing (decreasing) stock price identified the article as positive (negative). After the model had been trained, new articles were compared based on their similarity to the aligned ones. The higher the similarity, the higher the probability that the market will react. Gidofalvi (2001) used a naive Bayesian classifier for the prediction of the stock market; however, as he noted himself, the predictive power is not promising. Fung et al. (2005) have already used the more common support vector machine classifier. It is interesting that the authors recommended the inclusion of all articles in the analysis in 2002, due to the risk of violating the EMH. However, three years later this recommendation was ignored, and only some articles enter the training process. An explanation for this change in mind is missing.

The Wall Street Journal, as one of the significant information providers in the equity market, was used by Tetlock (2007) to demonstrate that negative wording and outlook influences trading behaviour. He used a column which summarized the previous trading day. In the author's opinion, the articles in the column can be used as sentiment indicators. The results suggest that negative content in the column leads to downward pressure on prices, with a revision to fundamentals afterwards. An increased trading volume can be seen as a side effect of the negative sentiment. The author notes that it is not clear whether the information in the newspaper has an amplifying effect or clearly reflects the expectations of the investors. However, the results are consistent with behavioural finance theories and the tendency to overreact to negative information. In 2008 Tetlock, Saar-Tsechansky and Macskassy picked up this idea and extended the analysis towards full articles related to S&P 500 companies. As before they have been able to prove that negative wording forecasts low firm earnings. Furthermore, they observed an under-reaction of the stock price to new information. The authors point out

that the qualitative analysis of language enables researchers to reveal new information about the company's fundamentals and in addition, the directional impact of multiple events can be studied at once, whereas other studies suffer from the limited number of events. General statements about patterns can, therefore, be made much more efficiently.

In a later work, Tetlock (2011) focused on general news articles and their impact on the equity market. According to the efficient market hypothesis, new information is immediately incorporated in the prices of stocks. He used a wide-ranging dataset with news articles about publicly traded companies and showed that news providers reuse specific information in short periods of days over and over again. The author measured the staleness of this information and found that individual investors trade more aggressively when the news is stale. This is even more observable when stocks are dominated by individual investors rather than institutional investors. Those results are similar to the findings of Lee et al. (1991). Tetlock (2011) used a comparative measure to estimate the similarity of unique words in the articles. The higher the similarity, the staler the news. He found that returns do not overreact to stale news, but the return of the day of the stale news does negatively predict the return of the following week. The author notes that his analysis excludes other economic drivers that might have influenced the behaviour of traders as well. Still, it seems that there is a negative amplifying process which pushes investors to overreaction when they are mirrored with the same information over a particular time.

Lee and Timmons (2007) also believe that the stock market is influenced by news articles. Based on the assumption that investors read the publicly available news, it is important to increase this field of research. The authors picked up the thought of Fung et al. (2005) and developed a similar text classification system, which categorized news articles with the help of a reference list of companies. Their results show that a passive trading strategy can be outperformed with their system. They tested the more straightforward bag of words approaches against the maximum entropy classifier. The research reached its limit by analysing too much data at once in terms of memory capacity. Nonetheless the maximum entropy classifier, which analysis one or more paragraphs of the news articles, seems to be superior in terms of prediction.

Since there is no consensus in the literature regarding the best way for sentiment extraction, Schumaker and Chen (2009) compared three commonly used methods: a bag of words, noun phrases and named entities. To submit evidence, the authors tried to use those three methods to predict stock prices in combination with a support vector machine (*SVM*) derivative, which

was introduced by Fung et al. (2002). Their study showed that textual based stock price prediction, with either one of these three used methods, is superior in comparison to a linear regression approach. Among the three different methods, however, named entities performed best. The reason for this can be found in the fact that articles are represented in a minimally way, due to the transformation through the algorithm. The authors see further improvement in their research when they narrowed their developed program down to just a few industry groups instead of focusing on all S&P 500 companies.

Another source of company information is earning announcements. Sadique et al. (2008) used those to reveal whether the tone in earning announcements has any impact on the returns of the company. Additionally, they used the financial news coverage of those announcements to get a better picture of the impact. The authors define tone as the ratio of positive and negative words. Similar to the work of Tetlock (2011) they used the pre-specified Harvard IV-4 psychological dictionary to define positive and negative words. Their analysis revealed that positive tone decreased the volatility and increased the returns of the stock, whereas the negative wording leads to a mirroring result.

Similar to the results of Godbole et al. (2007), O'Hare et al. (2009) decided to focus on financially related blogs. Their motivation is based on the fact that blogs do show more exact sentiment than news articles. In comparison to other studies, the authors decided to combine topic modelling and polarity classification. Blog posts are usually related to more than one company, and documents, therefore, show sentiment shifts where one company is favoured, and the other is not. The advantage of combining both methods is that the linked sentiment can be directly extracted from the texts.

Duric and Song (2012) presented a more theoretical overview of possible applications of NLP. They focused on topic modelling and the disadvantages for sentiment analysis. Similar to the previous scholars the authors put lexica and their composition into the centre of sentiment analysis. According to them, there are multiple ways to construct lexica. Research has shown that seed-based lexica that are extended by topic related terms might be a superior solution. This confirms the achieved results of Lavrenko et al. (2000).

All the above-mentioned examples have one thing in common. The analysis that the individual researchers perform aims at two things. First, analysis of the whole article, and second, categorization of the articles based on their sentiment into positive, negative or neutral. Nasukawa and Yi (2003) criticized this method due to the fact that traditional natural language processing achievements are going to be lost. The sole focus on the classification of words

ignores the relationship between them. The sentiment is usually not expressed as a whole but towards specific objects. The authors recommend that NLP operators should recollect previous knowledge and focus on text structures rather than simple quantification of words. The study of O'Hare et al. (2009) can, therefore, be seen to be in line with the criticism of Nasukawa and Yi (2003).

Gabrilovich and Markovitch (2009) criticize as well the performed analysis with a sole focus on polarity classification. If researchers focus only on parts of documents such as the headline, the actual understanding of the meaning is not guaranteed. However, this understanding is needed for the interpretation of topics and results. The authors propose using real-world lexica, such as Wikipedia, to improve understanding. Comparing parts of documents against Wikipedia entries results in an in-depth understanding of related topics.

More generally, Loughran and McDonald (2014) have argued that the increasing amount of textual analysis in the financial world requires a better understanding of the documents. Understanding of critical financial documents could be increased when the readability of those documents is improved. As a measure of readability, the authors criticize the single focus on the Fog Index, which has been increasingly used in the literature. The measure is insufficient for the financial world due to its construction. The Fog Index aims at sentence length and word complexity. Since 10-K filings are dominated by multisyllabic words, which are easily understood, the index suggests lower readability. Loughran and McDonald recommend researchers use the file size of the 10-K filing documents instead, where larger files stand for lower readability.

SUMMARY

It is quite difficult to grab NLP and its subcategories, topic modelling and sentiment analysis, within an increasing body of literature. This short overview has tried to show that NLP has emerged from the first attempts of interaction between humans and machines. In recent decades, the increasing use of computers in our day-to-day life has brought significant improvements to the field.

Documents have been identified as a significant source of sentiment and opinion in multiple fields. The use of written information, however, has provided researchers with a variety of new questions. Maks and Vossen (2012) have recognized that word lexica are a crucial significant element in the correct interpretation of sentiment. Even though self-defined lexica seem to be

superior to predefined ones, researchers need to be aware of personal subjectivity. Nearly all elements regarding the interpretation of words and opinions are subject to personal biases. A good example can be seen in O'Hare et al. (2009), where the manual categorization of text documents was performed by multiple individuals. Even though the authors declared that the participants had been trained, individuals interpreted documents differently. Other researchers either prefer that such classifications are done by only one person, or they recommend automated classifications.

The variety of studies in the equity market show that sentiment shifts and trading behaviour changes are more likely with companies which are dominated by noise traders. For instance, Mian and Sankaraguruswamy (2012) confirm the results of Lee et al. (1990) who have analysed the closed-end fund puzzle. Small, young, highly volatile and non-dividend paying stocks are more exposed to sentiment shifts. Furthermore, and consistent with other results, investors do respond differently to the news. Larger responses can be expected if the bad (good) news fits with the current underlying market mood.

The sole focus on news articles is criticized by O'Hare et al. (2009), among others, due to the fact that journalistic articles can be interpreted as objective rather than subjective. The extraction of sentiment is therefore limited. One advantage is the impact factor, which can be assumed to be larger in comparison to other media sources such as blogs. News articles will reach a wider audience.

A topic which was excluded from nearly all of the given examples is whether companies influence their media coverage actively or not. This question was raised by Ahern and Sosyura (2014). It seems that companies could impact their media coverage more actively and could, therefore, influence the sentiment. The timing of releasing specific information to the public depends on the company's intentions. Different timing may result in different reactions. The authors analysed merger processes and found that, during the negotiation process, the media coverage increases. The results suggest that it is possible to publish biased information to influence the stock prices actively.

The criticism of Nasukawa and Yi (2003) seems justified by summarizing the presented research. Scholars seem to be one-sided when it comes to textual analysis. Yet, the authors do ignore the fact that many of the criticized operators do not have a traditional linguistic background. Therefore, the majority of them try to simplify the applications as much as possible for the specific field of interest.

Since the research field heavily relies on data, researchers should be aware of quality issues which may arise. Rajakumari (2014) classified four categories of data quality, where each category depends on different dimensions such as accuracy, completeness, consistency and timeliness. Researchers could benefit from higher quality data sources. Rajakumari recommends a quantitative quality check of the data to identify where weaknesses are present. The presented research concentrates on online information (articles or blogs). The judgement as to whether the quality of this information is satisfying has not been provided by all scholars. Tetlock (2007), for instance, entirely bases his studies on *The Wall Street Journal*, which can be assumed to be a top-quality information source with a satisfying coverage. This can be seen as an argument in favour of the use of this information source.

To summarize, it can be concluded that the literature provides evidence that written information carries enough sentiment to show that a correlation between market developments and media coverage is present. This result is in line with fundamental behavioural theories. It seems that negative news remains longer in people's minds than positive news. As an example, I would like to mention Carroll et al. (1994), who stated that the citation of the bad sentiment, which was not measured in articles, but was extracted from the Index of Consumer Sentiment, led to an economic slowdown. Based on this evidence, opinion mining and textual analysis are rightfully identified as a source of sentiment.

4.2.3 NLP ON THE REAL ESTATE MARKET

NLP and the developed methods have been adopted in the equity market with success. Since real estate is not as frequently traded as stocks, researchers tend to apply equity market theories first to the REIT market. Doran et al. (2010) have analysed the content of quarterly earnings conference calls of publicly traded REITs and linked the tone of the calls back to the stock prices. They applied the proposed technique by Tetlock (2007) and used a customized dictionary and the *Harvard Psychosocial Dictionary*. Via the use of General Inquirer, the authors were able to extract the sentiment of the calls. Their analysis revealed that the Q&A part of those calls contributes more to the sentiment than the introductory speech of a chairman. A positive tone between the management and the analyst offsets negative feedbacks from negative company announcements. The authors were able to confirm the results for the equity market provided by Sadique and Veeraraghavan (2008).

Sentiment analysis based on text mining has reached the residential real estate market. Soo (2015) applied natural language-based techniques to the real estate market quite early. Motivated by the same observation as Case et al. (2012) or Foote et al. (2012), Soo (2015) thinks that the financial crisis has been analysed with a sole focus on the fundamental issues. The exclusion of sentiment and opinions is difficult to understand given the behavioural finance knowledge to hand. The decision to focus on the housing market for her study is based on the fact that housing is more often traded by individuals and that sentiment shocks are more readily identified. The study examines all cities which are present in the Case-Shiller Home Price Index. Applying the method introduced by Tetlock (2007), Soo (2015) filtered the tone of the news articles to develop her underlying sentiment index. Similar to previous studies she used the Harvard IV-4 Dictionary and included customized terms. Based on her study, she was able to forecast the financial market downturn with a lead of two years. The author showed that sentiment in news articles influences the real estate market.

Walker (2014a) extended the application of NLP to the real estate market. Based on a more significant corpus of news articles regarding the U.K. housing market, the author looked at the financial crisis and the influence of opinions which have led to irrational decisions. Walker examined the sentiment of the market with the help of Diction, a software application, which uses a word lexicon to interpret the documents. According to the author, sentiment influences average house prices. Furthermore, the results reveal that the sentiment or optimism in the market declined one year ahead of the crisis.

Building upon those results and those of Soo (2015), Walker (2016) showed that media coverage and influence on the behaviour of stock traders are much more far-reaching than assumed. He used news articles related to the U.K. housing market to see whether stock traders who trade U.K. housing company stocks are influenced by the sentiment of the articles. He used a similar approach to Freybote (2016), who also used a different underlying stock market which is linked to the market of interest as a proxy for their analysis. The results reveal that stock prices are influenced by the sentiment of the traders who are influenced by the sentiment of the housing market. Walker (2016) paid attention to the fact that the news articles are not linked to the stock market in particular. This study shows that we are just beginning to understand which factors lead to specific changes in our behaviour. It seems that people who have stocks of companies in a particular industry pay attention to the whole industry rather than just the company itself.

4.2.4 NLP: METHODOLOGICAL DEVELOPMENT

Different methods based on lexicon categorization have been developed over recent decades. Finn (2011), for instance, was focusing on the microblogging service Twitter, he recommended that sentiment extraction from text documents should be based on the comparison of words against a labelled list. He developed his own list and compared it with other lexica. In conclusion, his list showed better results regarding sentiment extraction. Today, there are two lists provided by the author. Each word carries a score from –6 (negative) to 6 (positive). His list has been developed with the help of ANEW, SentiStrength, General Inquirer and Opinion Finder. The author used a seed of pre-defined tweets to compare the effectiveness of the new list against the other lists; 1,000 tweets have been labelled by humans via *Amazon*s Mechanical Turk (AMT). This excellent result might be caused by the fact that his own list was initially developed for Twitter, while the other lists have a different origin and haven't been adjusted by the author. In the remainder of this work, I will refer to this method as *AFINN*.

A different approach was taken in Hu and Liu (2004), the authors introduced an improved method for opinion mining for product reviews. Different to earlier studies the authors used a small list of words, which is topic independent and extended the list via the use of WordNet.[12] Using sentences as the unit of interest within the text, the focus was set on those sentences which include adjectives. Those words are used to described features and opinions. The words are categorized into positive and negative. Starting with their base list, the authors used the organization of WordNet in bipolar adjective structures (synonyms) and generated a more substantial list of words, which all have a similar meaning. Yet the authors draw the conclusion that the recommended method reaches its limitation when the texts fail to show a clear separation into positive and negative descriptions. This appears in free-formatted reviews. In the remainder of this work, I will refer to this method as *BING*, named after the Liu Bing, who has developed the lexicon.

Mohammad and Turney (2010) developed an emotional lexicon via using the same method and drawing back on the opinions of real people with the help of AMT. Not primarily interested in the sentiment of people but in the emotions, which are awoken by precise terms, the authors assigned a list of words to different feelings. The primary motivation for the development of this lexicon was the fact that, first, those lexica do not exist and, second, that terms can trigger certain emotions and therefore influence the reader. In the remainder of this work, I will refer

[12] WordNet is a lexical database.

to this method as *NRC*, named after the funding body of the project: The National Research Council of Canada.

Another application of NLP and opinion mining is the Stanford Natural Language Toolkit (Stanford CoreNLP) for JAVA. Unfortunately, this method has not been included in the R-package I have been using. I will therefore only mention the method at this point, in order to present a full picture of all main methods. Stanford CoreNLP has been used in Socher et al. (2013) and Manning et al. (2014). The authors applied a deep learning algorithm to the problem at hand and introduced a Sentiment *Tree*bank. According to the study of Manning et al. (2014), the authors show that the model is able to outperform any other model significantly. They also used sentences as their smallest unit of opinion. The words within each sentence are scored into five different sentiment classes (very positive to very negative). *Tree*bank is based on a reasonably large corpus which is annotated. In combination with Recursive Neural Tensor Network, the model is able to identify even negations. Therefore, the proposed method is superior in comparison to the bag of word approaches as used in Pang and Lee (2008).

SUMMARY

This literature review has revealed that opinion mining is based on the interpretation of the wording within the document. The classification of texts into positive, neutral or negative, or any other scale, is an essential aspect. Yet, people interpret things differently, and therefore the developed lexica differ in words they include.

Most of the lexica are topically related. Well-known examples are ANEW, General Inquirer, Opinion Finder, SentiWordNet as well as WordNet. It has further become clear that even though many studies rely on a bag of words approaches, with downsized part of speech (POS) elements (i.e. words), it is no longer the most appropriate method.

Socher et al. (2013) have shown that with an increase in the number of words per POS the sentiment is more likely to be non-neutral. Further, the ignorance of the word order, which is done in other studies, is from a cognitive and linguistic point of view unclear.

The usage of these different approaches leads to two problems. First, since all of them rely on computer algorithms, the user is forced to learn, at least to a minor extent, some coding. Second, a comparison between the different methods is difficult since the lexica have been developed for a specific topic.

4.3 THEORY

Real estate service agencies publish a variety of different market reports on a regular basis. A broad distinction between commercial, residential as well as other property types is usual. The majority of these market reports can be downloaded from websites. Yet, those reports do not represent the primary business field of the companies. They are used for information provision as well as advertisement. Companies present their expertise and their track record. The reader should, therefore, be aware that those market reports are biased in a two-fold way: first, how the information is presented and, second, which information is presented.

I will follow Walker (2016) who focused on the U.K. housing market and its media coverage in the *Financial Times*. He suggested that different authors may have different information about the market. This should also apply to market reports from different companies. As already stated, not all companies offer the same set of services and should, therefore, have different fields of expertise.

An issue which arises from the usage of market reports is that the reports are relatively infrequently published in comparison to news articles. Walker (2016) points out that infrequent trade in real estate creates a gap which is not covered by the reports or by any other media, such as news. So, we face two lagged actions which might cause problems in the analysis: an infrequently traded commercial real estate market and infrequent media coverage. Yet, this infrequency is a characteristic of the market and a similar style of information coverage may be suitable.

The main concern regarding this gap can be found in the fact that the sentiment might change during this unreported period, due to macroeconomic or political factors. Another issue, in comparison to other studies such as Walker (2016) or Kothari et al. (2009), is that my dataset is relatively small with less than 1,500 documents. However, since in this chapter I aim to give an overview of the different sentiment extraction methods, I assume that a smaller dataset can still provide some useful insight.

One could argue that market reports reflect the perception of the service agencies and that this perception is partially driven by the market sentiment and observable developments. I like to refer back to the introduction of behavioural finance (chapter 1.3) and the work of Katz (1957). I think it is fair to describe the market reports as one form of opinion leadership. The service agencies demonstrate their expertise through the collection, analysis, and publication of market data. In addition, the authors of the reports draw conclusions from the most recent

development and present their personal perception about the upcoming developments based on that. Market participants, who read these reports might follow the presented opinion and adjust their behaviour accordingly. This does lead to an amplification of the presented perception and relates later again to some form of market sentiment, which is expressed through transactions, yields, constructions or rents. One cannot set the perception and the market sentiment equal. Because the read information is processed differently by each individual. This is why I follow the hypothesis that market reports are read on a frequent basis and that they, therefore, should be able to influence the sentiment within the market.

Given that, the discussion can be extended to the question. whether it is the reaction of market participants or the reaction of appraisers that is being tested in the analysis. Due to the fact that the textual sentiment indicators are tested against three MSCI total return indices for the British market, this question is valid. The dependent variables are appraisal based and each surveyor is influenced by the information they consume. At the same time, the literature review in the introductory chapter (1.3.1) has shown that surveyors are influenced by their behaviour and different biases as well. Among others, clients actively influence the valuation from time to time.

However, since the sentiment is based on the information extracted from market reports, which are essentially summarizing the most recent market developments. And given the fact that the valuations are done from various surveyors, I am convinced that it is the reaction of the market, which is being tested. Each valuation considers assumptions about market development. These assumptions need to be formed by the surveyors based on different sets of information. One source might be the discussed market reports. The biases and perceptions of the individual surveyors, similar to the above-discussed authors of the market reports, should blend, because of the use of multiple valuations in order to form the index value.

Within a modelling framework, I assume that the sentiment should have a positive influence on the total return indices. An increase in the market sentiment should go in hand with an increase in the returns.

4.4 DATA DESCRIPTION

Two different datasets are used in this section. The first dataset uses three MSCI total return indices for the British market. The first index is based on all property types in the U.K. In a second analysis the focus will be set on the U.K. office market and finally the London city office market will be analysed. All three series are given on a quarterly level ranging from 2005Q1 to 2016Q4. For comparison reasons and for robustness checks I also draw on the previously constructed indirect sentiment measures from chapter 3.

The second dataset is represented by market reports. NLP uses text documents and transforms them into quantifiable data. For the data collection, I used either an online '*grabbing* tool' (*GetThemAll* – a Google Chrome application) or downloaded the reports manually from the websites, which was done in three sessions, in February and April 2015 and one year later in April 2016.

I tried to present a full picture of all service providing companies on the commercial real estate market in the U.K. Therefore, I have tried to collect market reports from all larger service agencies. The data collection has resulted in a small text corpus of market reports from BNP Paribas Real Estate (133 documents), Cushman and Wakefield (143), CBRE (77), Colliers (176), DTZ (684), Jones Lang LaSalle (139), Knight Frank (355) and Savills (487).

Table 4:1 - Overview of all collected market reports

	BNPPRE	C&W	CBRE	Colliers	DTZ	JLL	KF	Savills	Total
Industrial	8	9	3		37	24	23	5	109
Office	93	36	8	39	338	46	48	88	696
Other	7	48	28	6	73	11	16	48	237
Residential	9	7	17	34	9	32	225	145	478
Retail	16	21	9	13	88	14	12	35	208
CRE		4	8	79	108	1	6	66	272
Investment		13			23			28	64
Magazine		2							2
Politics		3							3
Capital Markets			3				1		4
Student Housing			1				3	5	9
Caravan Park				1					1
Care Homes				4			2		6
E-Tailing					1			1	2
German open-ended funds (GOEF)					6				6
Nursing Homes					1				1
Hotel						1	2	10	13
Index						8	5		13
Olympics						1			1
Tech						1			1
Health Care							4		4
Retirement Housing							1		1
Rural							6	25	31
Survey							1		1
Development								31	31
Total	133	143	77	176	684	139	355	487	2,194

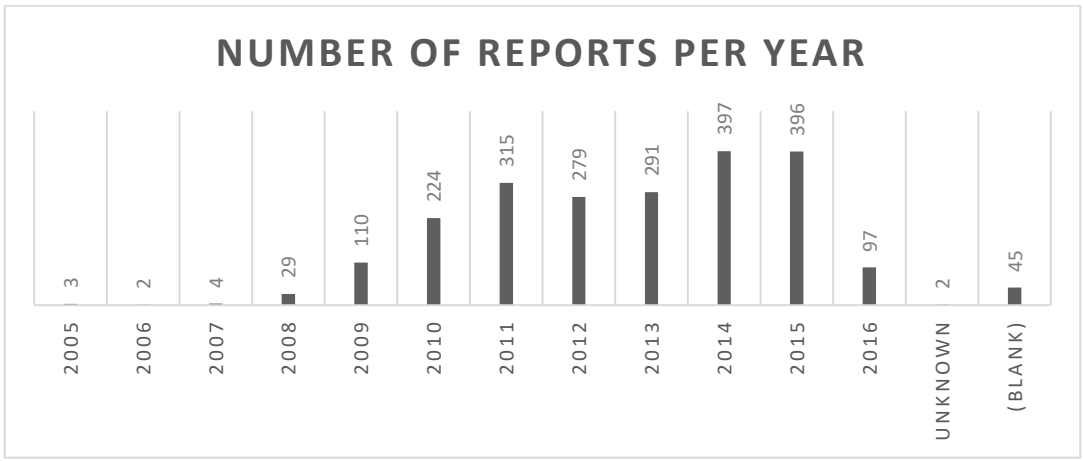
Note 4.1: The table shows all market reports by the company and with the corresponding property type or topic of the report.

Some websites did offer to preselect the property type and the region. However, as Table 4:1 illustrates, residential reports were collected for all companies as well. Those reports will be excluded from the analysis. The companies seem to delete older reports from time to time. Therefore, a regular data collection might be necessary to build up a significant corpus.

Since I aim to develop an index, the time component of the reports is essential. Unfortunately, I encountered several issues during the sorting process. Reports are published at different frequencies and cover different periods (month, quarter, year or season); it was, therefore, difficult to sort those reports accordingly. The majority of reports gave a specific description. Still, the documents are very infrequent, and at the least, the company-specific indices will suffer from missing information. The data has been sorted into quarters, in order to

generate a sufficient amount of reports for the analysis as well as allowing to compare the newly constructed sentiment measures, with those from chapter 3.

Figure 4:1 - Number of market reports per year



Note 4.2: The figure illustrates the amount of market reports per year. There are several reports, which have not been assigned to a year, due to the lack of information. For the first 4 years (2005 - 2008) only a small number of articles have been available.

Figure 4:1 illustrates the distribution of all collected market reports. The dataset reaches back until 2005 when only a few reports are available. Since 2009 the dataset is much more consistent with more than 100 reports per year.

A similar issue arose when I sorted the documents according to their region and their property type. As Table 4:1 shows, the variety of different categories is remarkable. The category "other" includes even more document types which I was unable to sort. Knight Frank, for instance, has even published a hunting lifestyle magazine, which can be seen as off-topic. Some categories such as CRE (commercial real estate), capital markets or investment cover multiple property types at once.

As stated earlier, the companies publish location-specific reports, e.g. for London or Manchester. Yet some reports cover multiple regions at once, such as the whole of the U.K. or the South East. London-specific reports are published by nearly all companies. This suits the purpose of this study since London is nationally, as well as internationally, a vital property market.

Given the above-described issues, I ran a set of four different analyses for each of the four lexicon methods on a quarterly base between 2005q1 and 2014q4 (40 quarters).

Table 4:2 - Overview of the planned analysis

Analysis	1	2	3
Market	U.K.	U.K.	London
Property type	Overall CRE	Office	Office
Company	All	All	All
Number of documents	897	619	150
Included categories	Capital Markets, CRE, Investment, Office	Office	Office

Note 4.3: The table shows the three planned analyses. Analysis 1, will use the largest share of reports and will also look at the broadest market. Analysis 2 will focus on the office market with a slightly smaller corpus. Finally, corpus 3 will look at the London office market with 150 market reports.

With each of the different sets, the number of documents within the underlying corpus decreases (Table 4:2). In total, I ran three different analysis on the commercial real estate market. One concerning the U.K. commercial market; one regarding the U.K. office market and third, one which is looking at the London office market.

The first analysis uses a more focused corpus, where only obvious commercial real estate market reports have been included. A total of 897 documents were used. This number differs severely from the overall collected number of reports. However, the mixture of several topics would only lead to a noisy corpus, which would reduce the overall explanatory power of the textual sentiment indicators.

The second analysis uses only documents which deal with the office market. This reduces the number of documents down to 619. The advantage of this focused corpus is that the office market is fully covered and that noise produced by other property types does not play any role.

Given the available data for the London market and due to the fact that roughly 200 documents (including residential and none office reports) share London as a frequent topic, the analysis is also performed on the London office market. The office specific corpus is the smallest one, with only 150 documents. I did not expect the textual sentiment indicators to perform very well since the indices are based on a small number of documents.

Table 4:3 illustrates the summary of statistics.

Table 4:3 - Summary of statistics: NLP

Label	Obs	Mean	Std. Dev.	Min	Max
IDP Total return index all properties	45	1.686	3.966	-12.958	9.992
IDP Total return index all offices	45	2.094	4.158	-12.671	8.243
IDP Total return index all offices in the City of London	45	1.382	4.747	-14.802	15.686
Interest rate	45	2.022	2.130	0.500	5.750
Macroeconomic Sentiment	43	-0.799	0.289	-1.460	-0.307
Office Sentiment	42	0.609	0.725	-1.000	2.120
Google Trends	44	0.208	0.633	-1.238	1.039
Textual Sentiment Indicator: All distinct commercial related market reports for the U.K. (*BING*)	36	0.086	0.063	-0.050	0.190
Textual Sentiment Indicator: All office related market reports for the U.K. (*BING*)	35	0.108	0.080	-0.050	0.260
Textual Sentiment Indicator: All office related market reports for London (*BING*)	33	0.116	0.085	-0.200	0.250
Textual Sentiment Indicator: All distinct commercial related market reports for the U.K. (*AFINN*)	36	0.396	0.140	0.160	0.640
Textual Sentiment Indicator: All office related market reports for the U.K. (*AFINN*)	35	0.381	0.135	0.160	0.680
Textual Sentiment Indicator: All office related market reports for London (*AFINN*)	33	0.373	0.153	-0.100	0.720
Textual Sentiment Indicator: All distinct commercial related market reports for the U.K. (*NRC*)	36	0.692	0.114	0.490	0.970
Textual Sentiment Indicator: All office related market reports for the U.K. (*NRC*)	35	0.719	0.123	0.530	1.020
Textual Sentiment Indicator: All office related market reports for London (*NRC*)	33	0.643	0.127	0.420	0.960
Textual Sentiment Indicator: All distinct commercial related market reports for the U.K. (*TM*)	36	65.533	57.077	26.230	270.000
Textual Sentiment Indicator: All office related market reports for the U.K. (*TM*)	35	65.598	51.465	25.860	270.000
Textual Sentiment Indicator: All office related market reports for London (*TM*)	33	69.965	26.844	21.860	131.000

Note 4.4: The table presents the summary of statistics for the Natural Language Processing dataset.

Table 4:4 illustrates the Augmented Dickey-Fuller test for stationarity. It can be seen, that the different sentiment components do not have a unit root. The dependent variables on the other hand needed some statistical modification. To detrend the series I used the logarithm and in addition, I needed to take the first difference to reach stationarity.

Table 4:4 - Augmented Dickey-Fuller Test

Variable	Test statistics	1% critical value	5% critical value	10% critical value	Obs.
IPD Total return index all property types U.K.*	-0.553	-2.441	-1.691	-1.307	40
IPD Total return index all property types U.K. (1st difference of log)*	-2.563	-2.445	-1.692	-1.308	37
IPD Total return index all offices U.K.*	-0.288	-2.441	-1.691	-1.307	40
IPD Total return index all offices U.K. (1st difference of log)*	-2.714	-2.445	-1.692	-1.308	37
IPD Total return index all offices London City*	-0.162	-2.441	-1.691	-1.307	40
IPD Total return index all offices London City (1st difference of log)*	-2.94	-2.445	-1.692	-1.308	39
Macroeconomic sentiment*	-1.748	-2.462	-1.699	-1.311	35
Office sentiment*	-3.076	-2.467	-1.701	-1.313	34
ZGT*	-2.508	-2.479	-1.706	-1.315	32
AFINN: All distinct commercial related market reports for the U.K. (Standardized) **	-4.305	-4.316	-3.572	-3.223	32
AFINN: All office related market reports for the U.K. (Standardized)	-1638.021	-3.702	-2.98	-2.622	32
AFINN: All office related market reports for London (Standardized)	-3.141	-3.709	-2.983	-2.623	31
BING: All distinct commercial related market reports for the U.K. (Standardized)	-3.023	-3.702	-2.98	-2.622	32
BING: All office related market reports for the U.K. (Standardized)	-2649.739	-3.702	-2.98	-2.622	32
BING: All office related market reports for London (Standardized)	-3.577	-3.709	-2.983	-2.623	31
NRC: All distinct commercial related market reports for the U.K. (Standardized)	-4.958	-3.702	-2.98	-2.622	32
NRC: All office related market reports for the U.K. (Standardized)	-4.918	-3.709	-2.983	-2.623	31
NRC: All office related market reports for London (Standardized)	-4.373	-3.709	-2.983	-2.623	31
TM: All distinct commercial related market reports for the U.K. (Standardized)	-3.226	-3.709	-2.983	-2.623	31
TM: All office related market reports for the U.K. (Standardized)	-3.314	-3.709	-2.983	-2.623	31
TM: All office related market reports for London (Standardized)**	-3.188	-4.362	-3.592	-3.235	27

* consideration of a drift

** consideration of a trend

*Note 4.5: The table illustrates the results of the Augmented Dickey-Fuller Test. The first panel illustrates the results for the three dependent variables. I needed to take the first difference of the logged time series to make the variables stationary. Other series had either a drift (indicated by *) or a trend (indicated by **) component.*

4.5 EMPIRICAL FRAMEWORK

This section is divided into four parts. The first part will present the autoregressive model, which I use for comparison reasons. Second, I introduce the standard terminology of NLP and text processing. Then, the different steps of pre-processing, especially text cleaning, will be described. Finally, the four different methods and their idiosyncrasies are presented.

It is of importance to draw a line between sentiment and opinions at this stage. Even though both terms are used as synonyms, the sentiment is just one element of the opinion itself. Following the methodology of Liu (2012), an opinion is characterized by five elements: the target entity (e_j), one aspect of the entity (a_{jk}), the sentiment (so_{ijkl}) of the opinion from the opinion holder (h_i) towards the feature of the entity at a certain time (t_l):

$$opinion\ (e_j, a_{jk}, so_{ijkl}, h_i, t_l)$$

Liu (2012) points out that opinion without any target is useless. I have followed the general methodology of Liu (2012) and like to extract the sentiment towards either the U.K. or London commercial real estate or office market. The opinion holders in this context are the report providing service agencies, based on the sample of usable reports identified in Table 4:2.

4.5.1 BASE MODEL

To compare the quality of the constructed indicators, I ran a simple autoregressive model, $AR(1)$, on three different IPD (Investment Property Databank) portfolio total return indices.

To compare the overall performance of the commercial real estate corpus, I use the total return index for all properties ($ukipdqtrall$). To have a closer look at the office specific reports for the whole U.K., I use the total return index for office properties ($ukipdqtrof$). For those market reports which are centred around the London office market, I utilize the total return index for office properties in the City of London ($ukipqtrofc$).

The different sentiment indicators from the market reports will be added to the base models successively. A similar model has been presented in Tsolacos (2006), where the author tested the effect of interest rates and GDP on the IPD measure.

$$\Delta lTRET\_ALL_t = \alpha + \beta_1 \Delta lTRET\_ALL_{t-1} + \beta_2 \nabla \text{CRE\_TEXT\_SENT}_{t-i} + \varepsilon_t \qquad \textit{Equation 4:1}$$

where $TRET\_ALL_t$ is the IPD all properties total return index on a quarterly level. As suggested by the model, the indices are logged and the first differences are taken, as indicated by Δl. Through the introduction of the lagged dependent variable ($\Delta lTRET\_ALL_{t-1}$) as an explanatory variable, market developments from the previous period are considered. $\text{CRE\_TEXT\_SENT}_{t-i}$ represents the four different textual sentiment indicators based on the commercial real estate corpus. The indicators have been standardized with a mean of 0 and a standard deviation of 1, as indicated by ∇.

$$\Delta lTRET\_OFF_t = \alpha + \beta_1 \Delta lTRET\_OFF_{t-1} + \beta_2 \nabla \text{OFF\_TEXT\_SENT}_{t-n} + \varepsilon_t \qquad \textit{Equation 4:2}$$

where $TRET\_OFF_t$ is the IPD all offices total return index on a quarterly level. $\text{OFF\_TEXT\_SENT}_{t-i}$ represents the four different textual sentiment indicators based on the overall office related real estate corpus. The remaining model components are unchanged.

$$\Delta lTRET\_OFF\_CITY_t = \alpha + \beta_1 \Delta lTRET\_OFF_{t-1} + \beta_2 \nabla \text{LONDON\_OFF\_TEXT\_SENT}_{t-n} + \varepsilon_t \qquad \textit{Equation 4:3}$$

where $TRET\_OFF\_CITY_t$ is the IPD all offices total return index for the City of London on a quarterly level. $LONDON\_\text{OFF\_TEXT\_SENT}_{t-i}$ represents the four different textual sentiment indicators based on the London office document corpus. The remaining model components are unchanged.

The optimal number of lags has been estimated by reducing the Akaike Information Criteria (AIC).

The chosen model might seem too simple in order to prove my assumption to be correct. One could argue, that the models lag several control variables such as macroeconomic factors

(i.e. GDP or the interest rate). However, by solely focusing on the textual indicator its magnitude and influence on the dependent variable becomes clearer.

4.5.2 TERMINOLOGY

4.5.2.1 CORPUS

The text corpus is the base for any textual analysis [Bird et al. (2009)]. It consists of a body of text documents, where each corpus is directed towards one specific topic. In this case, it is the commercial property market in the U.K. and London.

4.5.2.2 TOKENIZATION

Tokenization describes the process where the corpus is separated into words and/or sentences. Both methods have been used over recent decades. Some scholars such as Socher (2013) believe that sentence tokenization is superior in comparison since the order of words carries essential information. Furthermore, it has been shown that longer *ngram* units (multiple words), such as sentences, are more often non-neutral regarding the sentiment they carry.

Both methods need to have a clear text body. European languages use both white spaces to separate words and punctuation to separate sentences from each other. An algorithm is able to identify these signs and split the corpus accordingly. Palmer (2010) illustrates a range of difficulties regarding language separation.

University of Reading *Example 4:1*

Example 4:1, for instance, illustrates the point that separation into individual words would destroy the logical unit.

I need to tell you that Mr. Heinig has cancelled the meeting. *Example 4:2*

Example 4:2 shows the issue of separating sentences. The algorithm needs to identify that the abbreviation “Mr.” is not the end of the sentence. This plays an essential role in the part-of-speech tagging process. The algorithm should, therefore, be able to distinguish between the different punctuations of the English language (period, comma and semicolon) and should further know the structure and usage of these.

4.5.2.3 NORMALIZATION AND STEMMING

This step is done to simplify the corpus. Morphological normalization reduces a variety of words to their stem (Example 4:3). All those words carry the same information. However, they only differ because of linguistic reasons. The stemming process does not remove any additional information but decreases the total number of words within the corpus.

houses become ***house*** or ***drinking*** becomes ***drink*** *Example 4:3*

Savoy and Gaussier (2010) itemize a range of different examples. Example 4:3 only illustrates the stemming process for suffixes, but prefixes are also removed. The R - package uses the Porter Stemmer, as introduced by Porter (1980).

Other essential steps, which are summarized under the label of normalization, transform the text to lower cases and remove stop-words, such as “a” and “the”.

Also, numbers are removed from the corpus. However, this might influence the significance of the analysis, since market reports usually consist of a variety of numbers. Nevertheless, it is not clear how to interpret the numbers during automated sentiment analysis.

4.5.2.4 LEMMA

A similar process to morphological normalization is lemmatizing. This method has its origin in the field of text translation and lexical analysis. A lemma dictionary is used, and words are translated into words with similar meaning. Therefore, lemma could be equalized with synonyms [Bird (2009)].

4.5.3 PRE-PROCESSING: EXAMPLE

In the following, each of the above-described steps is illustrated in one document. I use the Cushman & Wakefield – Market beat Office Snapshot Q1 2014 as an example. Each market report was downloaded from the website as a PDF. Before I could apply the text analysis, the documents were transformed into text files (.txt). I used R for this initial demonstration. The corpus cleaning process uses the provided functions of the Text Mining package by Feinerer et al. (2008).

After transforming the document into a text file, the total number of counted characters (including numbers) is 5,654. In a first step *additional* white space is removed from the document with the result of removing one character (5,653).

Excerpt from the market report:

```
Q1 2014 OVERVIEW The occupational performance of the UK office market
is improving, with strong competition for Grade A space supported by
robust demand and limited supply. Investor demand has also sustained,
and not just for Central London prime property as has been the trend in
recent years. However, the availability of quality assets is in tight
supply in most markets, which is inhibiting investment activity and
driving demand towards second-tier markets.
```

Next, all cases will be transformed into lowercase:

```
q1 2014 overview the occupational performance of the uk office market
is improving, with strong competition for grade a space supported by
robust demand and limited supply. investor demand has also sustained,
and not just for central london prime property as has been the trend in
recent years. however, the availability of quality assets is in tight
supply in most markets, which is inhibiting investment activity and
driving demand towards second-tier markets.
```

Numbers are removed:

```
q overview the occupational performance of the uk office market is
improving, with strong competition for grade a space supported by robust
```

```
demand and limited supply. investor demand has also sustained, and not
just for central london prime property as has been the trend in recent
years. however, the availability of quality assets is in tight supply
in most markets, which is inhibiting investment activity and driving
demand towards second-tier markets.
```

Now all stop words will be removed from the corpus. Stop words are used to link different parts of the text and guarantee a better understanding and flow of the text. However, they are not necessary for sentiment extraction:

```
q overview occupational performance uk office market improving,
strong competition grade space supported robust demand limited
supply. investor demand also sustained, just central london prime
property trend recent years. however, availability quality assets
tight supply markets, inhibiting investment activity driving demand
towards second-tier markets.
```

Next, the words are transformed into their root:

```
q overview occup perform uk offic market improving, strong
competit grade space support robust demand limit supply. investor
demand also sustained, just central london prime properti trend
recent years. however, avail qualiti asset tight suppli markets,
inhibit invest activ drive demand toward second-ti markets.
```

It can be seen that the applied word stemmer within the *TM* package (Porter Stemmer 1980) does not transform all words correctly. For instance, “sustained” is not transformed into “sustain”. After this pre-processing process, the total number of characters in the market report has been decreased to 4,052.

4.5.4 SENTIMENT EXTRACTION

AFINN, *BING*, *NRC* and *TM* can be run in R. I use the “*syuzhet*” package by Jockers (2016) for this analysis, since it summarizes the first three methods. The “*syuzhet*” package draws back on the “tm” package regarding the pre-processing of the corpora. *TM* or topic modelling has been widely used and a variety of plug-ins have been developed over the years. Among others, a sentiment specific plug-in is available, which is utilized for the analysis, as in the fourth method.

Besides the methods presented here, a variety of other methods are available. Some need a deeper understanding of other programming languages such as Python. A well-known

representative would be the Stanford CoreNLP application or the Natural Language Tool Kit (NLTK).

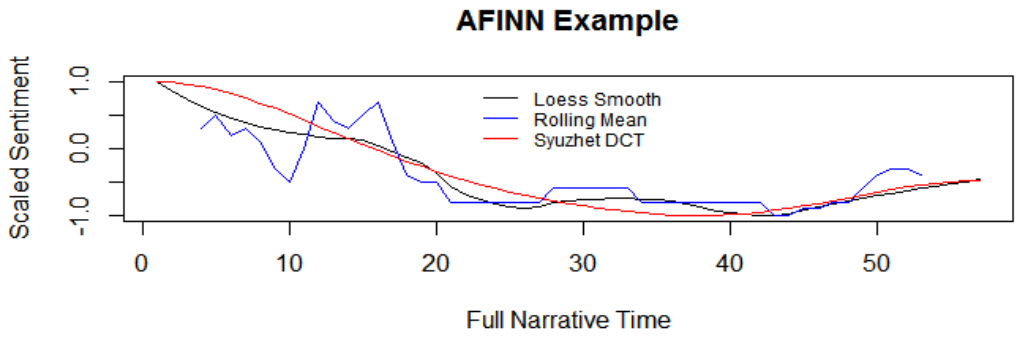
The chosen methods for this study rely on categorized word dictionaries, mainly sorted into positive and negative words. In the following, the individual methods and their specifics will be summarized.

4.5.4.1 AFINN

The second method is based on the work of Nielsen (2011). Similar to Liu et al. (2005) the author developed his own dictionary. One of the main reasons was that the Twitter Tweets he analysed showed a different wording than other texts. He collected a range of positive and negative words and scored them manually. This provided the author with higher accuracy since algorithms in many cases are a static structure.

Different to the previous method, the author scored the terms in a range between –5 and 5, which delivered a more detailed analysis. Nielsen (2011) finally ran a correlation analysis with his new dictionary against other methods (SentiStrength, Opinion Finder and the General Inquirer) and against labelled entities by humans (*Amazon*'s Mechanical Turk). The latter was used as a reference point. His method generated a higher positive Pearson correlation in comparison to the other three methods.

Figure 4:2 - AFINN example

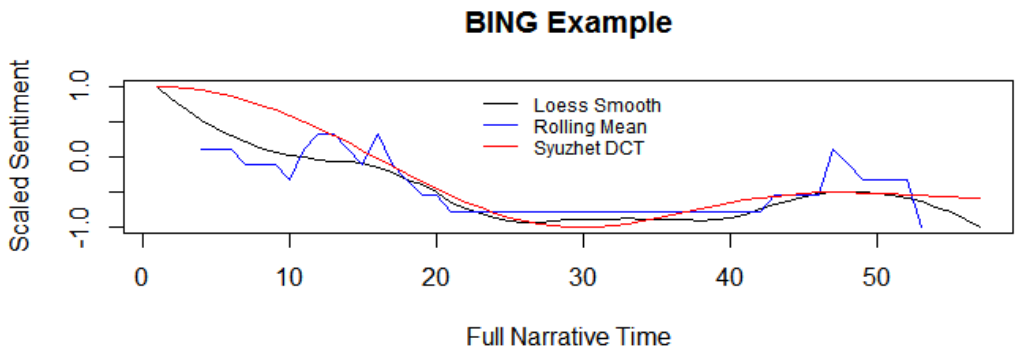


Note 4.6: The graph illustrates the sentiment within the example file: Cushman & Wakefield – Market beat Office Snapshot Q1 2014. I used the AFINN method to extract the sentiment from the file. The graph shows three different illustrations, a Loess Smooth graph (locally weighted scatterplot smoothing), the rolling mean of the positive and negative relations within each sentence, and the Syuzhet DCT (discrete cosine transformation). The sentiment has been scaled to a range from (–1) to 1.

4.5.4.2 BING

The first method is based on the work of Hu and Liu (2005) as well as Liu et al. (2005). As pointed out earlier, the authors were motivated to improve the reviewing process of products. Due to the vast amount of online product reviews it has become more difficult to read all reviews as a customer. The authors, therefore, developed a sentiment analysis which translates into a graphical visualization. The authors used the semantic meaning of words and grouped them into positive and negative categories. They used WordNet and a set of 30 words (positive and negative) as a starting point to develop their classified dictionary.

Figure 4:3 - BING example

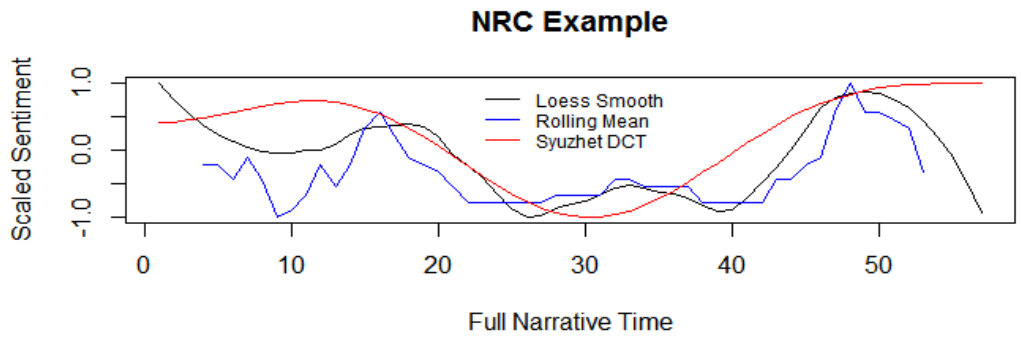


Note 4.7: The graph illustrates the sentiment within the example file: Cushman & Wakefield – Market beat Office Snapshot Q1 2014. I used the BING method to extract the sentiment from the document. The graph shows three different illustrations, a Loess Smooth graph (locally weighted scatterplot smoothing), the rolling mean of the positive and negative relations within each sentence, and the Syuzhet DCT (discrete cosine transformation). The sentiment has been scaled to a range from (–1) to 1.

4.5.4.3 NRC

A different approach was taken by Mohammad and Turney (2010). They identified a lack of lexica which measure emotions. Again, they drew on *Amazon*'s Mechanical Turk to categorize their entities. Different words create different emotions based on their context. Given the humanized categorization, the precision of their lexicon is satisfying. The *syuzhet* help file does not offer any insight as to which part of the word lexica from the *NRC* is used. Given the fact that I am able to measure the positive and negative words, I assume that the included lexica ignores the emotional sorted words for the sentiment extraction and refers to the positive and negative labelling of each word.

Figure 4:4 - NRC example

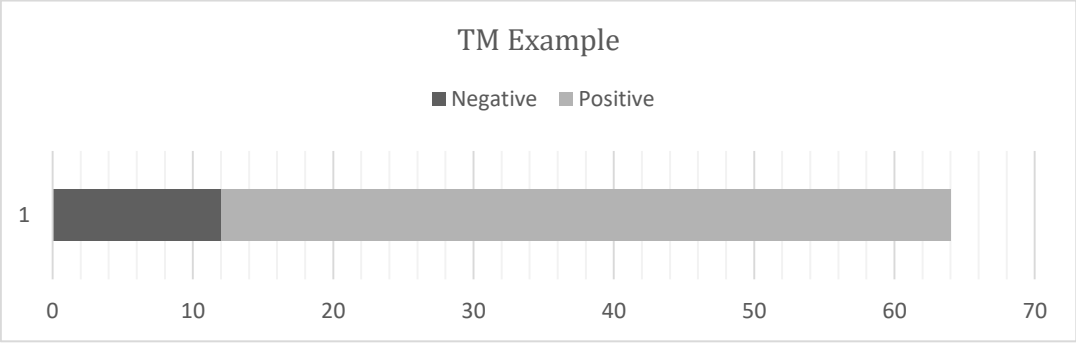


Note 4.8: The graph illustrates the sentiment within the example file: Cushman & Wakefield – Market beat Office Snapshot Q1 2014. I used the NRC method to extract the sentiment from the file. The graph shows three different illustrations, a Loess Smooth graph (locally weighted scatterplot smoothing), the rolling mean of the positive and negative relations within each sentence, and the Syuzhet DCT (discrete cosine transformation). The sentiment has been scaled to a range from (–1) to 1.

4.5.4.4 TOPIC MODELLING (*TM*)

The *TM* package and different plug-ins make the program a useful source for NLP. I apply the tm.lexicon.GeneralInquireR - package of Theussel. The package links the analysis to the Harvard General Inquirer Dictionary. This lexicon has been used in a variety of studies [Maynard and Bontcheva (2016); Kiritchenko and Mohammad (2016)] and can be seen as one of the more reliable sources in the NLP world. The lexica are organized in different categories and summarize four different sources. We assume that the *syuzhet* package draws on the positive and negative categorization within the Harvard IV-4 Dictionary.

Figure 4:5 - Topic modelling example



Note 4.9: The graph illustrates the sentiment within the example file: Cushman & Wakefield – Market beat Office Snapshot Q1 2014. I used the TM.Sentiment.Plugin to extract the sentiment from the file. Unfortunately, the sentiment results are not presented at a sentence level; only the overall scores for positive and negative words are given.

All four methods are based on word lexica. Table 4:5 illustrates the number of words, the separation into neutral, positive and negative words as well as the initial purpose. It has further become clear that in all four cases the number of negative words exceeds the number of positive words, which might indicate why negative word counts perform better since the underlying dictionaries are of a finer grade on this side.

Table 4:5 - Overview of the different lexicons

	AFINN		BING	NRC	TM
Name	AFINN-96	AFINN-111	Opinion Lexicon	EmoLex	General Inquirer: H4 and H4Lvd
Initial purpose	Twitter Tweets		Product reviews	Measuring of emotions	Multiple
Number of words	1468	2477	6788	14182	11787
Neutral	1	1	0	0	0
Positive	515	878	2005	2312	1915
Negative	964	1598	4783	3324	2291
Score	1 - 5		0 or 1	0 or 1	positive or negative

Note 4.10: The table illustrates the four different sentiment lexicon and their initial purpose.

4.6 RESULTS

In the following, I will present the results of the three different subcorpora. The dependent variable will be adjusted according to the focus of the corpora that has been used to construct the textual sentiment indicators.

4.6.1 AUTOREGRESSIVE MODEL

For the first test, the sentiment indicators are based on all market reports which have explicitly discussed the U.K. commercial real estate market. Only those from the collected articles that belong to the Capital Markets, CRE, Investment or Office category within England, Scotland, Wales and Northern Ireland have been considered. This has reduced the number of reports significantly.

A total of four sentiment indicators have been constructed. In all cases, the indicator represents the overall average from all sentiments per document. So, each indicator is based on the mean value of positive and negative words per document.

For the first analysis, I use the IPD Total Return Index for all properties as the dependent variable. Table 4:6 illustrates the results of the four textual sentiment indicators in the AR (1)

model. The base model which only relies on the lagged version of the dependent variable reaches an R-squared value of 0.586. The only independent variable is highly significant at a 1% level, while the constant is insignificant. The base model uses a total of 43 observations. Running the standard statistical tests, I encountered heteroscedasticity in the base model. Therefore, the reported errors are robust and control for the presence of heteroscedasticity.

Looking at the four textual sentiment indicators, only the *TM* indicator is able to produce a significant coefficient at the 1% level. Unexpectedly, the sign is negative. Meaning that an increase in the sentiment has a negative influence on the total return. Different to the base model all four textual sentiment models show highly significant independent variables. For the *TM* model, the R-squared value lies at 0.796, which is a significant improvement upon the base model. Even though the remaining models failed to produce significant sentiment coefficients, they also show significantly higher R-squared values, ranging between 0.689 (*BING*) and 0.712 (*NRC*). All textual sentiment indicators enter the autoregressive model with three lags. This number has been estimated by reducing the AIC.

For comparison reasons, I have further added the previously constructed sentiment indicators. It can be seen that only the macroeconomic sentiment measure is able to produce a significant sentiment coefficient at the 10% level. Again, the coefficient has a negative sign which is unexpected at this stage. The constant for all three models remains insignificant. These indicators have also entered the model with different lags. Comparing the R-squared values, both the macroeconomic (0.637) and the Google Trends measure (0.598) show a marginal improvement on the base model.

The second analysis tests those indicators which have been constructed with the help of all office market reports. As described before the number of reports has been dropped to 619. Table 4:7 illustrates the results of the autoregressive model. The dependent variable is now the IPD total return index for office properties. The overall results have been improved compared to the previous analysis. The coefficient of the independent variable in the base model is highly significant at the 1% level. The constant, however, remains insignificant. The R-squared value is now 0.636.

Looking at the textual sentiment indicators, the results for the four coefficients have been improved. The coefficients of the *AFINN* and the *BING* model are highly significant at the 1% level. The *TM* model shows a significance at the 5% level. Only the latter model has all components significant. Comparing the R-squared values the *TM* model once more produced the highest value at 0.833. Both the *AFINN* and the *BING* model have an R-squared value of

0.721. Similar to the above-presented results, all significant coefficients have a negative sign, which is somewhat surprising.

Again, the previously constructed sentiment indicators have been added. Different to the textual sentiment indicators no improvement upon the first analysis can be observed. Only the macroeconomic indicator is significant at the 5% level. The model reaches an R-squared value of 0.675, which when compared to the textual sentiment indicators is somewhat marginal in terms of improvement.

The last point, which is worth mentioning, is the fact that all indicators enter the model with at least one lag. This seems reasonable since the market reports are a description of the past. Most of them are further not published immediately but more than a quarter behind the described market development.

Table 4:6 - Result for the AR (1) model: overall commercial document corpus

VARIABLES	Labels	Base Model	Macroeconomic Sentiment	Office Sentiment	Google Trends	AFINN	BING	NRC	TM
dlipdtrall = L,	IPD total return all properties (first differences of the log)	0.761***	0.625***	0.743***	0.716***	0.607***	0.614***	0.610***	0.542***
		[0.142]	[0.111]	[0.146]	[0.126]	[0.059]	[0.063]	[0.063]	[0.041]
macroeconomic_sentiment = L,	Macroeconomic Sentiment (lagged)		-0.042*						
			[0.021]						
office_sentiment = L,	Office Sentiment (lagged)			-0.008					
				[0.006]					
ZGT = L,	Google Trends (lagged)				-0.013				
					[0.008]				
z_*AFINN*_uk_mix = L,	*AFINN* (lagged)					0.001			
						[0.004]			
z_*BING*_uk_mix = L,	*BING* (lagged)						0		
							[0.000]		
z_*NRC*_uk_mix = L,	*NRC* (lagged)							0.005	
								[0.005]	
z_tm_net_uk_mix = L,	*TM* (lagged)								-0.011***
									[0.004]
Constant		0.003	-0.028	0.007	0.007	0.010***	0.010***	0.010***	0.013***
		[0.006]	[0.019]	[0.007]	[0.006]	[0.003]	[0.003]	[0.003]	[0.003]
Observations		43	40	39	37	33	34	33	30
Number of lags		-	3	5	3	3	3	3	3
AIC		-187.797	-175.345	-167.144	-156.810	-170.677	-176.892	-173.130	-167.936
BIC		-184.275	-170.287	-162.154	-151.977	-166.188	-172.313	-168.641	-163.733
R-squared		0.586	0.637	0.586	0.598	0.69	0.689	0.712	0.796
Adjusted R-squared		0.576	0.617	0.563	0.575	0.669	0.669	0.693	0.781
F-Statistic		28.82	17.84	13.65	16.19	52.88	49.24	54.7	89.21
Degrees of freedom		41	37	36	34	30	31	30	27

Robust standard errors in brackets *** p<0.01, ** p<0.05, * p<0.1

Note 4.11: The table shows the result of the overall commercial real estate corpus for the U.K. The dependent variable is the IPD total return index for all properties. The textual sentiment indicators use 897 market reports including the following categories: capital markets, CRE, investment and office.

Table 4:7 - Result for the AR (1) model: all office related market reports

VARIABLES	Labels	Base Model	Macroeconomic Sentiment	Office Sentiment	Google Trends	AFINN	BING	NRC	TM
dlipdtroff = L,	IPD total return all offices (first differences of the log)	0.795***	0.728***	0.746***	0.756***	0.765***	0.764***	0.766***	0.622***
		[0.132]	[0.111]	[0.120]	[0.119]	[0.137]	[0.137]	[0.138]	[0.039]
macroeconomic_sentiment = L,	Macroeconomic Sentiment (lagged)		-0.041**						
			[0.018]						
office_sentiment = L,	Office Sentiment (lagged)			0.005					
				[0.007]					
ZGT = L,	Google Trends (lagged)				-0.013				
					[0.008]				
z_*AFINN*_uk_office = L,	*AFINN* (lagged)					-0.014***			
						[0.001]			
z_*BING*_uk_office = L,	*BING* (lagged)						-0.014***		
							[0.001]		
z_*NRC*_uk_office = L,	*NRC* (lagged)							0.002	
								[0.004]	
z_tm_net_uk_office = L,	*TM* (lagged)								-0.010**
									[0.004]
Constant		0.004	-0.027	0.001	0.008	0.004	0.005	0.007	0.013***
		[0.006]	[0.017]	[0.008]	[0.006]	[0.006]	[0.006]	[0.006]	[0.003]
Observations		43	40	39	37	35	35	34	30
Number of lags		-	2	1	3	1	1	1	3
AIC		-187.914	-176.671	-165.411	-157.372	-156.831	-156.806	-151.359	-169.696
BIC		-184.391	-171.604	-160.420	-152.539	-152.164	-152.139	-146.780	-165.492
R-squared		0.636	0.691	0.639	0.651	0.721	0.721	0.679	0.833
Adjusted R-squared		0.627	0.675	0.619	0.631	0.704	0.703	0.659	0.820
F-Statistic		36.520	22.180	20.740	20.360	916.800	929.400	15.950	125.000
Degrees of Freedom		41	37	36	34	32	32	31	27

Robust standard errors in brackets *** p<0.01, ** p<0.05, * p<0.1

Note 4.12: The table shows the result for the office corpus for the U.K. The dependent variable is the IPD total return index for all offices. The textual sentiment indicators use 619 market reports.

The last autoregressive model uses the IPD total return index for all offices in the City of London. The results have once more slightly improved upon the first two models, although the base model still does not provide a significant constant and the R-squared value has improved up to 0.64. The independent variable remains highly significant.

Looking at the textual sentiment indicators again the *AFINN*, the *BING* and the *TM* model have significant sentiment coefficients. This time, however, no model produces a significant constant. The *AFINN* and the *BING* model with their highly significant sentiment coefficients outperform the *TM* and the remaining models. The *AFINN* model reaches an R-squared value of 0.744 followed by the *BING* model (0.742). The contribution of the *TM* model is this time a bit smaller, and the goodness of fit measure only reaches a value of 0.713. Despite the inadequate model specification, the *NRC* model also outperforms the base model. This time the AFINN and the BING model reveal the expected sign, while the remaining models still have a negative impact on the dependent variable.

Comparing the indirect sentiment measures to the textual sentiment measures, it can be seen that this time two of the three models are significant. The macroeconomic sentiment model has a highly significant coefficient at the 1% level and reaches an R-squared value of 0.714. The second significant model (5% level) is the Google Trends model with an R-squared of 0.662.

While before all sentiment induced models entered the model with at least one lag, this time both the *AFINN* and the *BING* model show the smallest AIC value with no lag.

Table 4:8 - Result for the AR (1) model: all office related market reports for London

VARIABLES	Labels	Base Model	Macroeconomic Sentiment	Office Sentiment	Google Trends	AFINN	BING	NRC	TM
dltret_office_city = L,	IPD total return all offices in the City of London (first differences of the log)	0.799***	0.710***	0.756***	0.787***	0.558***	0.655***	0.764***	0.748***
		[0.135]	[0.110]	[0.129]	[0.133]	[0.128]	[0.121]	[0.144]	[0.139]
macroeconomic_sentiment = L,	Macroeconomic Sentiment (lagged)		-0.052***						
			[0.017]						
office_sentiment = L,	Office Sentiment (lagged)			0.005					
				[0.007]					
ZGT = L,	Google Trends (lagged)				-0.017**				
					[0.007]				
z_*AFINN*_london_office	AFINN					0.022**			
						[0.009]			
z_*BING*_london_office	BING						0.019***		
							[0.007]		
z_*NRC*_london_office = L,	*NRC* (lagged)							-0.001	
								[0.003]	
z_tm_net_london_office = L,	*TM* (lagged)								0.007**
									[0.004]
Constant		0.004	-0.035**	0.001	0.01	0.006	0.005	0.008	0.008
		[0.007]	[0.017]	[0.009]	[0.007]	[0.005]	[0.006]	[0.007]	[0.006]
Observations		43	40	39	37	33	33	32	32
Number of lags		-	2	1	2	0	0	1	1
AIC		-180.857	-172.772	-159.258	-151.256	-142.649	-142.387	-137.271	-139.644
BIC		-177.335	-167.705	-154.268	-146.243	-138.159	-137.888	-132.874	-135.247
R-squared		0.640	0.714	0.644	0.662	0.744	0.742	0.691	0.713
Adjusted R-squared		0.631	0.699	0.624	0.643	0.727	0.725	0.670	0.693
F-Statistic		35.150	22.290	18.520	25.610	22.400	26.470	15.040	16.270
Degrees of Freedom		41	37	36	34	30	30	29	29

Robust standard errors in brackets *** p<0.01, ** p<0.05, * p<0.1

Note 4.13: The table shows the result for the office corpus for London. The dependent variable is the IPD total return index for all offices in the City of London. The textual sentiment indicators are based on 150 market reports.

To conclude, the analysis of the three different sub corpora has shown that the focus on a more precise topic within the documents has helped to improve the statistical values. All sentiment induced models were able to outperform the base model. While for the first two the best results have been achieved by using the *TM* model, the last has shown further improvement of the other models: *AFINN* and *BING*. The *NRC* model, on the other hand, did not produce any significant coefficient. The comparison of the different sentiment indicators has further shown that those indicators, which are based on indirect sentiment measures, fail to outperform the textual sentiment indicators. This result was not entirely expected but does provide an interesting observation.

4.6.2 ROBUSTNESS CHECK

I will provide one robustness check, where the quality of the textual sentiment indicators should be evaluated. I draw on the comparison between the constructed sentiment measures and the direct sentiment measures, provided by RICS.

Table 4:9 - Robustness check: correlation analysis (RICS)

	(1) U.K. RICS property survey: sales & rental levels, London, next qtr	(2) U.K. RICS property survey: sales & rental levels, London, next qtr	(3) U.K. RICS survey: office sales & rent levels, London, next qtr nadj
AFINN	0.683	0.098	0.602
BING	0.120	0.097	0.513
NRC	-0.102	-0.136	0.321
TM	0.245	0.390	-0.124

Note 4.14: The table illustrates the correlation analysis between the 4 constructed sentiment measures and the direct sentiment measures for the London property market (U.K. RICS surveys). Each column does use a different set of lexical sentiment measures. The first column is using the overall sentiment measures based on the full corpus. The second column does use the CRE sentiment, and the last column is using the London office specific sentiment measures for the analysis.

Table 4:9 illustrates the correlation analysis between the four corresponding textual sentiment indicators and the three adequate direct sentiment measures. In column 1, the textual sentiment indicators refer to the commercial real estate market report corpus. Column 2 refers to the all-office section and column 3 to the office section for the London market. It can be seen that the highest correlation is achieved by the *AFINN* indicator (0.683) for the all properties survey measure. In the second column, a weak correlation between the *TM* indicator (0.390) and the London office measure can be observed. For the last column, the correlation

results improve again, and both the *AFINN* and the *BING* model show a moderate correlation with the RICS measure.

Even though these correlations are not as good as they have been for the sentiment proxies, it can be stated that the textual sentiment indicators resemble some market sentiment for the British market.

4.7 CONCLUSION

A variety of sentiment measures have been applied to the equity and the real estate market. Studies have emphasized that direct measures are superior in comparison to indirect measures. Yet, it needs to be asked how the opinion, expressed in a survey, has been formed. A survey represents a summary of a range of opinions, which have been manifested before.

Three sources for professionals to build an opinion have been identified: experience, information exchange with co-workers and information collection. Where the first two are difficult to measure, this chapter has used market reports for sentiment extraction. The four applied methods have different origins and therefore differ in their ability to express the underlying sentiment. One goal of this study was to provide a smooth and reproducible method for sentiment extraction. The method used in Walker (2016) would require access to the program DICTION. R and the R packages are free of charge, which guarantees reproduction.

In this chapter, I have illustrated that sentiment can be extracted with the help of natural language processing. While the utilization of macroeconomic factors seems more logical for real estate market participants, the collection, modification and construction on the other side, are more complicated in comparison to the use of text documents.

Service agencies use market reports to summarize market development and to give an outlook for the future, so they incorporate both back and forward-looking elements. Further, market reports can be seen as one of the significant information providing documents in the market. The application of different word lexica has shown that, given the underlying nature of the lexica, sentiment can be extracted. However, not all lexica provide similar results. While both the *AFINN* and the *BING* models have proven to be flexible, the *NRC* model did not provide satisfactory results. The *TM* model, which uses one of the major lexica in the field, outperformed the other three models in two of the three cases. Surprising is the fact, that the coefficients showed some sign flipping. While the significant textual sentiment indicators in the first two

tries remained negative, the AFINN and the BING model showed the expected positive relationship with the dependent variable. Reasons for this inconsistency are not clear. It seems counterintuitive that the measured sentiment should have a negative influence on the dependent variable.

The results show that the collection of documents and the restructuring of the corpus is of essential importance. However, I have generated satisfying results even with a rather small corpus of documents. This confirms the initial hypothesis that market reports carry underlying market sentiment. Market participants should not ignore the opinion which is expressed in the documents. The significant textual sentiment indicators were able to improve all the base models throughout the entire study. This can be seen as a confirmation of the previously presented theory in Liu (2012), where sentiment needs to be linked to a specific topic.

During the work, multiple obstacles have been identified. The primary concern regards the size of the corpus. According to Keller and Lapata (2003), size matters. In my dataset, some years are only represented by a deficient number of market reports. Other studies such as Kothari et al. (2009) or Walker (2016) used 10,000 to 100,000 documents. Also, the different slices of the analysis have lowered the number of reports down to less than 200. I am aware that this gives a biased result.

Another limitation of this study can be found in the methodology itself. The removal of numbers is generally seen as a necessary step during the pre-processing of the corpus. However, numbers are an essential element of market reports and experienced market participants are able to read and interpret their meaning.

Different to the methodology in chapter 3.4.2 I have taken the textual sentiment indicators as they are. One could argue, that they are still influenced by other known or observable factors and they could be stripped from those influences by orthogonalizing them as well. Future research will show, how textual sentiment indicators might benefit from such a statistical modification.

Nevertheless, this chapter has proven, and this can be seen as a central implication for the industry, that service agencies have the power to influence the market with the wording they use in the documents. The aggregation of quantified market reports is able to mirror the market sentiment for the U.K. CRE market.

5 MACHINE LEARNING APPLICATION

5.1 INTRODUCTION

In the previous chapters, I have used macroeconomic and textual sentiment proxies to extract market sentiment. In both cases, it has become apparent that the consideration of sentiment is able to provide a substantial insight into the market and that base models benefit from adding the sentiment.

While the macroeconomic sentiment proxies might be more understandable for market participants, they rely on a variety of collected variables and partially on a sophisticated way of construction. Textual sentiment indicators, on the other hand, rely on only one set of variables, and, with a minimal understanding of coding, sentiment can be extracted.

The advantage of this rather innovative data source lies in the improvement of the frequency. While most of the macroeconomic variables use backwards-looking information and are further published after market development, text documents can be seen as closer to the market. The dataset used in Chapter 4 has only a minor part of this advantage since the market reports are also published one to three months after specific developments.

However, these initial results from the previous chapter have encouraged me to proceed. In this chapter, a new dataset of more than 100,000 news articles concerning the U.K. real estate market, between 1 January 2004 and 31 December 2015 (144 months), will be analysed with a range of supervised learning algorithms and word lists. The extracted sentiment, for a selected number of methods, will enter in a second step a probit model, to examine how the textual sentiment might be able to improve predictions.

Scholars and market participants rely on a range of sentiment proxies, which improve models to some extent; however, the search for a universal proxy remains unsuccessful. The studies which rely on proxies are bound to either the specific property type or to a specific region. Surveys, for instance, which are assumed to be superior in comparison to other methods, have been used in a range of different studies [Vohra and Teraiya (2013), Kauer and Moreira (2016), Pang et al. (2002), Dave et al. (2003), Fang and Chen (2013), Nguyen et al. (2015) and Abbasi et al. (2008)], yet they either differ regarding their structure or do not cover all markets at once.

Furthermore, the criticism can be made that surveys are published after the sentiment has been formed. So, they are only reflecting the atmosphere at the time when the survey was created and therefore do not represent the sentiment at the time of publication. However, the reader might be influenced, and the publication can cause a multiplier effect on the market. Coming back to the three obvious factors of how a decision maker is influenced, I assume that the information stored in written documents carries a stronger and more essential sentiment since it can be measured instantaneously.

Three methods for sentiment extraction are commonly applied: a lexicon-based approach supervised learning and an unsupervised learning approach. The lexicon-based approach relies heavily on the ability to choose positively and negatively assessed words. The analysis of the corpus is then based on a term frequency of positive and negative words. Problems with this approach are that the number of words, as well as the correct labelling of the words on the topic related context, influence the results significantly. Some words might have a definite meaning in one topic but not in another. According to Medhat et al. (2014), the main issue concerns the process of how the lexicon is generated since in many cases topics are ignored, and the lexicon is just generated by synonyms and antonyms.

The other two approaches belong to the field of machine learning. Schapire and Freund (2012) define machine learning as the study of automatic methods for future predictions based on past observations. Both unsupervised and supervised approaches can be used for classification problems. The unsupervised approach is not yet widely used. In general, a computer algorithm tries to analyse an unstructured dataset by identifying patterns.

Supervised learning approaches, which are at the centre of this part of the thesis, also belong to the methods of pattern recognition. They use different mathematical and statistical theories to analyse an unknown dataset based on a known labelled dataset. In this chapter, nine different widely used algorithms for sentiment extraction will be tested and compared.

The supervised approach requires pre-knowledge of a share of a corpus. Typically, the corpus is divided into a training and a test dataset. The training share should be labelled so that an algorithm is able to learn based on the attached categories. The trained model will afterwards predict the categories for the test share. The central issue is the process of labelling the training documents. Other studies have either used an already labelled corpus [*Amazon* reviews: Dave et al. (2003); Hu and Liu (2004)] for their analysis or labelled the corpus manually [Chen et al. (2016); O'Keefe et al. (2013)]. To my knowledge, a labelled corpus for the real estate market, and especially for the U.K. market, is not available. To label a corpus manually, one either needs

to read a fair share of the corpus by oneself or one needs to utilize other methods such as *Amazon* Mechanical Turk. Besides the financial aspect of the latter, both personal biases and topic familiarity influence the outcome of the labelling process [Kauer and Moreira (2016)]. Another problem is the number of documents within the corpus. For instance, this chapter uses more than 100,000 news entities.

I assume that *Amazon* book reviews are a suitable source for people's opinions. The advantage of these book reviews is that each text is labelled with a rating (1 to 5 stars) by the authors. I further assume that people who read real estate related books might (a) be professionals or at least familiar with real estate as a topic and (b) might use a topic related *jargon* which is also reflected in the news entities. I have, therefore, crawled[13] more than 200,000 real estate related book reviews from www.amazon.co.uk and used them as the required training dataset.

The supervised learning algorithms will be trained on different sets of the *Amazon* book reviews. Those trained classifiers will then be used to extract the sentiment from the news articles. Based on the average score of the news entities and their aggregation on a monthly level, a sentiment score will be estimated.

The results of this study suggest that *Amazon* book reviews provide only marginal information to the probit models. They are outperformed continuously by the more straightforward lexicon approaches. Reasons for this can be found in the fact that book reviews are foreign topics to the real estate market. These latter sentiment measures are able to provide enough predictability. Robustness checks illustrate a close resemblance of the measures to survey-based sentiment measures and to the previously used sentiment measures. Nonetheless, classifying news articles, with the help of word lists, and then training supervised learning classifiers on this new training corpus, has produced excellent results, where lexicon measures are outperformed by the supervised learning measures (5.6.2).

The remainder of this chapter is structured as follows. In the next section, I will point out the relevance of sentiment analysis for the real estate market and will summarize the most recent research on NLP and text mining. Afterwards, the theoretical approach will be illustrated and the datasets, as well as the methodology, will be described. Finally, I will present a comprehensive analysis. I conclude with a summary of the key findings.

[13] This is an automatic process where specific information is extracted from single or multiple websites.

5.2 LITERATURE REVIEW

More conventional approaches as taken by Baker and Wurgler (2006) or others rely on sentiment proxies or survey data. Scholars have criticized these approaches for several reasons, but mainly because proxies do not measure sentiment in the first place, and surveys do not reflect the sentiment at the time when they are published.

More recent approaches allow for quantification of text documents. News articles, social media data or product/movie reviews [He (2012), Chen et al. (2016)], incorporate sentiment and opinions. Both scholars and market participants have identified this kind of document as a suitable source. However, there is no agreement yet as to which method or approach is suitable to generate overall satisfying results.

A more significant number of studies have analysed sentiment with regards to the stock market. Some have relied on conventional methods such as sentiment proxies [Frugier (2016); Liang (2016); Aissia (2016); Labidi et al. (2016)].

Other financial industries such as the banking sector have also applied textual analysis for credit risk or asset valuation [Smales (2016); Tsai et al. (2016)]. Smales (2016) used the Thomson Reuters News Analytics tool for his analysis, a dataset which incorporates documents which have been labelled by former market participants. This underlines the comments of other scholars which stress that the manual labelling process is more successful when background knowledge is given. The author concludes that negative articles have a stronger effect on the markets. A similar conclusion was reached by Tsai et al. (2016). They also focused on the count of negative words within the articles, because they would have a stronger influence on the reader. The authors are in line with Tetlock et al. (2008) who comment that positive word counts ignore the occurrence of negation and would, therefore, draw a blurred picture. One explanation for our tendency toward more negative words can be found in Soroka and McAdams (2015). These authors showed that even though people would prefer more positive news, they tend to focus on negative articles and headlines, somewhat subconsciously. From the perspective of a news agency, more negative news or headlines increase the readership, while positive headlines on a cover page, for instance, cause the opposite. Soroka and McAdams further point out that negative events are more likely to be remembered and we may have a stronger interest in these events because we may have to adjust to a new environment.

Scholars have identified that, based on Liu's (2012) terminology, a sentiment which is directed towards a topic has more value than a generally expressed sentiment. In this context,

Liu (2012) stressed that opinion without a target is one without use. Based on this Saif et al. (2016), Lin, Y. H. C. et al. (2012) and Lin, C. et al. (2012) used a common sentiment topic method for their analysis. They identified that, within one text, multiple topics can be discussed and that the overall sentiment might differ from topic to topic. Lin, C. et al. (2012) further state that labelled classifiers often fail to produce satisfying results within a new category. More flexible algorithms should be able to extract sentiment from multiple topics at once without any adjustments. The authors used a rather small corpus of just 2,000 documents. They further point out that an index based on social media data is correlated with socio-economic indicators and consumer confidence.

Not surprising but worth noting is the observation by Lin, C. et al. (2012) that documents seem to be influenced by previous documents dealing with the same topic. This can be compared with a wave effect, where one major event causes multiple and ever-increasing waves. Nguyen et al. (2015) applied a similar approach, used a common sentiment topic method and created a model to run predictions for stock movements based on social media data. They point out that social media data is characterized by short texts with misspelling and grammatical issues, which need to be addressed in the text pre-processing stage. It has become clear that Twitter data is noisy and not as useful as direct news sources. To overcome the grammatical issues within social media data, Fersini et al. (2016) focused on emoticons as a source for sentiment; this ignores the wording and makes the interpretation one-sided since emoticons can also be used in a sarcastic manner.

Also driven by the issues which arise through the labelling process, Kauer and Moreira (2016) developed a new method SABIR (sentiment analysis based on Information retrieval) and compared their results to the *SVM*, *MAXENT* and Naive Bayes algorithm. They used a corpus of Twitter tweets for their analysis and generated superior results.

He and Zhou (2011) point out that annotated corpora with sentiment classification lack the chance of portability across different domains and they, therefore, favour a self-learning approach. Different from other scholars He and Zhou (2011) move the focus to the feature level away from the entity level. Also, Fernández-Gavilanes et al. (2016) propose an unsupervised method for the sentiment analysis of online data. Again, they hope to automate the labelling process. The authors have the opinion that individual words matter more than their relationship to each other. However, their methods only achieve comparable results in relation to other methods.

The advantage of a weakly supervised or even an unsupervised learning approach can be found in the fact that the whole process of labelling becomes unnecessary. However, an unsupervised learning approach seems impossible to implement due to the range of multiple topics within a news article. And even the suggested unsupervised method by Fernandez-Gavilanes et al. (2016) can be seen as a weakly supervised approach since they apply the lexicon approach, where words have been labelled beforehand.

A different approach is taken by O'Keefe et al. (2013) who focus on quotes from the text documents. These quotes are directed towards a feature and might give a better indication of the sentiment. In general, an author tries to present the topic to a broader audience and is, therefore, addressing multiple opinions at once, which subsequently leads to a smoothing effect of the individual sentiments at the end. In their study, the authors limit the number of annotators to three to guarantee consistency during the labelling process. They used the Fleiss kappa measure to illustrate how similar the results of the different annotators are. In Chen et al. (2016) it is also underlined that the annotation of a single user is worth more than that of multiple users. This summarizes the general issue when it comes to manual labelling of the text corpus and controls for the fact that only the social biases of one person influence the labels.

5.3 DATA DESCRIPTION

In this section, I will briefly describe the four different datasets. The first three datasets have been used for the construction of the textual sentiment indicators. The *MSCI* dataset, on the other hand, was used to apply the textual sentiment indicators in a simple probit model.

5.3.1 NEWS ARTICLES: TEST DATASET

The main dataset has been collected via ProQuest U.K. News & Newspapers. The service provides access to a variety of U.K. based newspapers and was formerly known as U.K. Newsstand.

During the time when I collected the data, the site was reorganized, and some of my search parameters were changed. The U.K. News stream is now merged in the European News stream. The original search was performed on a monthly basis, due to the fact that the website only displays approximately 1,000 articles per search. I discovered that the search function of the

tool, which allows the pre-filtering of articles, is highly sensitive to the search terms. The data was collected with these parameters: English language, newspapers in the U.K. and full text search; and with these search terms: Savills, BNPPRE, DTZ, Jones Lang LaSalle, JLL, Cushman & Wakefield, office property, retail property, commercial property market, REIT, real estate investment trust and London. A total of 118,842 articles were displayed. However, during the crawling process, only 109,103 articles were downloaded. Reasons for this are unknown. Each entity is identifiable by date, publisher, title and full text of the article.

Even though the search terms aimed to be focused on the real estate market, this original corpus seems to be noisy. I have therefore decided to construct several sub-corpora, which in my opinion reduce the noise within the corpus. This follows the idea of other researchers that the sentiment should be analysed towards a specific feature. The search parameters also collected a number of housing-related articles; therefore, the first sub-corpus excludes all housing articles. I removed all articles which included the words: residential, housing, home, apartment or house; this reduced the number of articles from 109,103 to 62,266. However, this general exclusion might have excluded articles which discussed the broader real estate market. Nevertheless, I assume that the smaller corpus does focus more on the commercial real estate market.

A second sub-corpus was created and only includes articles with the word *London* (74,266 articles). That does not mean that all articles solely analyse the London real estate market; however, the chances are high that the property market of the city is at the centre of the discussion.

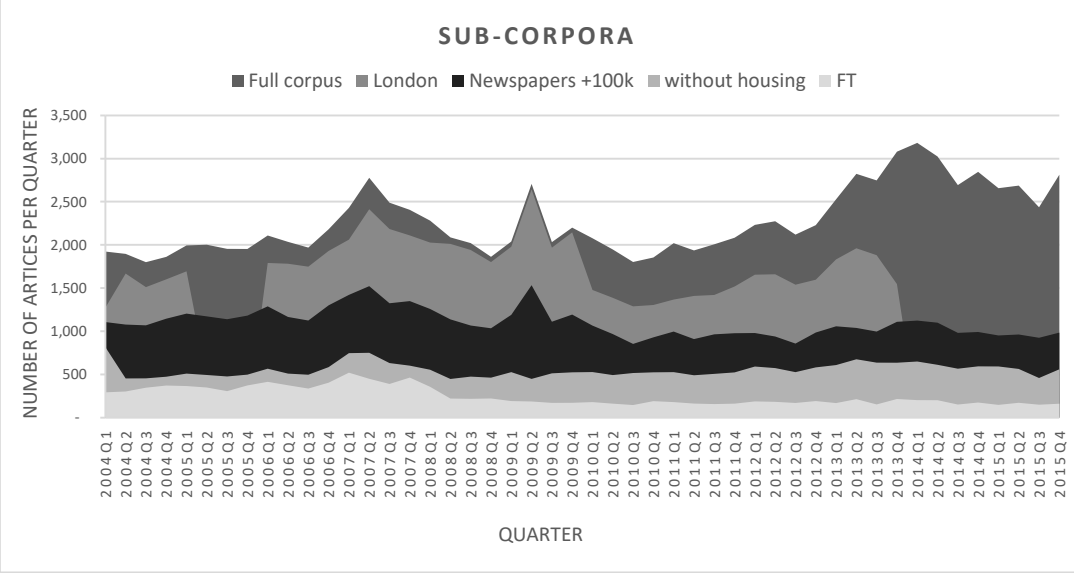
I am further interested in whether newspapers with a circulation above 100,000 papers per day might be able to influence the market in a stronger way; so, the third sub-corpus only includes: *The Daily Mail,* the *Daily Record, The Evening Standard, The Financial Times, The Daily Mirror, The Daily Telegraph, The Guardian, The Sun* and *The Times* (52,954 articles).

Since I want to examine the commercial real estate market and how market participants are influenced by news, I decided further to separate out all *Financial Times* entities. I believe that professionals are more likely to read the *Financial Times* than other newspapers (11,948 articles).

Figure 5:1 illustrates the distribution of articles per sub-corpora per quarter. It can be seen that the overall corpus shows some variation. The corpus regarding London shows that there were no observations at the end of 2005 and after 2013. It can be further seen that in 2007q2

and 2009q2 the number of articles peaked. This does not hold for all corpora but is influenced by the coverage of the financial crisis. Interesting is that after 2007 the number of articles in the *Financial Times* corpus dropped and remained steady with roughly 180 articles per month.

Figure 5:1 - Number of articles per sub-corpora per quarter



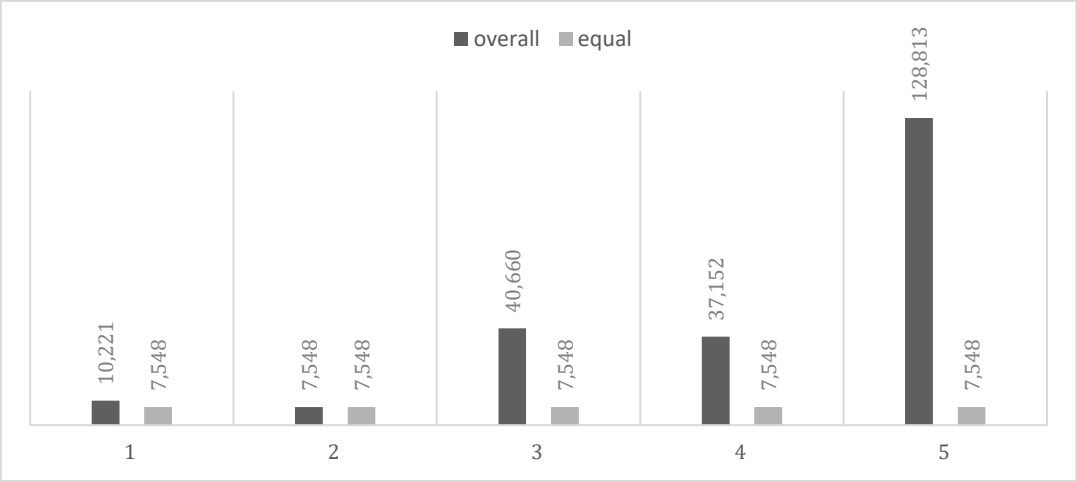
Note 5.1: The figure illustrates the overall distribution of all articles per quarter.

5.3.2 AMAZON DATA: TRAINING DATASET

The second dataset of this study consists of *Amazon* real estate related book reviews. I have crawled over 224,000 book reviews from around 5,800 books from www.amazon.co.uk.[14] Each book review has a rating between one (negative) and five stars (positive). The books were selected with the following search terms: real estate investment, property investment, real estate economics, real estate finance, real estate private equity, real estate valuation, property management, property valuation, property finance and real estate investment trust. Taking a closer look at the data two things become clear. The crawling process downloaded a range of reviews for books which are not related to real estate (e.g. intellectual property) and second, people tend to rate the books in a more positive way. In the collected dataset 57% of all reviews are rated with five stars. Figure 5:2 illustrates that more people give neutral to positive ratings than negative ones.

[14] The website was accessed on 12 March 2018.

Figure 5:2 - Rating of the reviews



Note 5.2: The figure illustrates the distribution of the Amazon Book review ratings for the overall and the equalized corpus. The overall corpus reveals a tendency towards the positive rating (5 stars). The equalized corpus does use 7,548 reviews for all categories based on the smallest number of reviews within one category (category 2).

This creates another issue for the labelling process. A model that is trained on this dataset would tend to the neutral or positive category. I have therefore created a smaller training dataset, which is equally distributed over the five categories with 37,740 reviews (7,548 reviews per category).

The literature seems to favour three categories (positive, neutral and negative) rather than five. I have created, based on the initial corpora, another two training corpora with just three sorting options. Over the training and testing process, the machine learning algorithms seem to perform better when they encounter fewer sorting options. In total, I have created four training corpora based on the *Amazon* book reviews (Table 5:1).

Table 5:1 - Amazon book review training corpus

Training corpus	Number of book reviews	Rating
1	224,394	1-5 stars
2	37,740	1-5 stars
3	224,394	positive - neutral - negative
4	37,740	positive - neutral - negative

Note 5.3: The table illustrates the four constructed training corpora. Corpus one and two use a five-category rating, while three and four rely on three categories.

Transforming the star ratings (Table 5:2) into the categorical ratings leads to a shift in the categories. One and two stars are transformed into negatives, three stars become neutral, and the remaining two have been assigned to the positive category.

Table 5:2 - Transformation of the categories

All reviews					
Stars	1	2	3	4	5
Reviews	10,221	7,548	40,660	37,152	128,813
Categories	Negative		Neutral	Positive	
Reviews	17,769		40,660	165,965	

An equal number of reviews					
Stars	1	2	3	4	5
Reviews	7,548	7,548	7,548	7,548	7,548
Categories	Negative		Neutral	Positive	
Reviews	15,096		7,548	15,096	

Note 5.4: The table above presents another detailed explanation of how the training corpora are constructed. It can be seen, that the overall corpus has a stronger tendency towards the positive side since three times as much reviews belong to the positive (category 4 and 5) category.

The newly assigned categories have shifted more weight to the negative and positive category in the equal training corpus and much more weight to the positive category in the training corpus which uses all reviews.

The last issues that arise from the *Amazon* book reviews are the labels themselves. On a linguistic and subjective level, some of the given ratings seem out of order. However, I wanted to interfere as little as possible in this initial trial. Yet, it seems debatable that “ok” as a stand-alone comment has a rating range from 1 to 5. The same applies to “awesome” or “excellent”:

subjectively I would rate books with these comments in the upper scale. Table 5:3 illustrates some of the issues I encountered within the reviews.

Table 5:3 - Example of the range of rantings

Comment	Rating range
Good	1, 3, 5
Awesome	3 - 5
Excellent	3 - 5
Ok	1 - 5

Note 5.5: The table illustrates several examples from the book reviews. It can be seen, that these words have been used to describe the quality of the book. However, there is no consistency in the corresponding rating.

5.3.3 FINANCIAL TIMES DATA

Given these facts and the rather weak model results, which will be discussed in section 5.6.2, I decided to create another corpus only using *Financial Times* entities. The reason for this is that the originally assumed similarity between the wording of book reviews and news articles is lower than expected. Since this corpus is not labelled, I am following Blum and Mitchell (1998); Nigam et al. (2000) and Liu et al. (2004) and use the lexical approach to label this training corpus before it enters the machine learning process (5.6.2). Another 55,872 articles were collected from ProQuest Newsstand. There is an overlap of 1.35% between the two corpora. The majority of the newly collected articles is not property related.

5.3.4 *MSCI* DATA

For the probit model, where I will test whether the textual sentiment indicators are able to predict turning points, the *MSCI* all property all asset and all office capital growth indices will be used (Table 5:4). Both will be modified into a binary or dichotomous variable with values of 0 and 1. One will represent those instances with negative growth. The *MSCI* data is available on a monthly level from January 2004 to February 2017, which provides in total 158 observations. According to the IPD Index Guide, “capital growth is calculated as the change in capital value, less any capital expenditure incurred, expressed as a percentage of capital employed over the period concerned”. Due to the fact, that no transactions, within the index-construction, are

considered[15], both series are essentially valuation driven. Reasons are, that the index should only reflect the actual market returns and should ignore unusual developments of the property which are caused by the individual management. This leads back to the discussion of chapter 4.3 and the question if the chosen dependent variable is suitable since it is not clear if the reaction of the market or the reaction of the appraisers is measured. As I have argued before, I assume that there is a fair chance that the blurring of multiple valuations, performed by different valuers should overcome this issue. Each valuation is based on assumptions taken from the market. These assumptions should be regularly corrected given new developments within the market.

Table 5:4 - Descriptive statistics for the dependent variable

Panel A - Binary Capital Growth series	All assets_all properties Jan2004 - Dec2015	All assets_office Jan2004 - Dec2015
Percentage of observations with negative growth	29.17%	26.39%
Obs.	144	144
Mean	0.292	0.264
Std. Dev.	0.456	0.442
Min	0	0
Max	1	1

Note 5.6: The table provides the descriptive statistics of the MSCI data set.

[15]Please refer to: https://www.msci.com/documents/1296102/1378010/Indexes+and+Benchmark+Methodology+Guide.pdf/bfbd2637-581d-411e-bd5f-34d0d2b6b9c1, accessed on 22.11.2018

5.4 EMPIRICAL FRAMEWORK

In general, the literature distinguishes between a lexical [Liu, Hu and Cheng (2005), Finn (2011) and Mohammad and Turney (2010)] and a machine learning approach [Maynard and Funk (2011), Muhammad, Wiratunga and Lothian (2016), He (2012)]. While the lexical approach has been widely used, some issues have been identified. First, it is crucial to select the right words within the right context for the word lists. Second, the amount of the words within the list are essential, since shorter lists might miss important words.

On the other hand, scientific issues need to be addressed. Some scholars have the belief that the order of words does not affect the sentiment within a document. Yet, sentiment extraction based on wordlists fails to detect negations or sarcasm, which are essential linguistic features. Scholars favour an n-gram approach or the analysis of the whole sentence. These issues do not exist with supervised machine learning approaches since the training documents are not analysed on a word or sentence level.

In this chapter, I use the R - package RTextTools by Jurka et al. (2012).[16] The package offers nine different algorithms: Support Vector Machine (*SVM*), Maximum Entropy (*MAXENT*), Stabilized Linear Discriminant Analysis (SLDA), Generalized Linear Model (*GLMENT*), Bootstrap Aggregation (BAGGING), Algorithm Enforcement (*BOOSTING*), Random Forrest (RF), Decision *TREE* (*TREE*) and Neural Net (*NNET*). Unfortunately, the Naive Bayes [17] and the Nearest Neighbour approach are not covered by the package.

In total, four different sets of classifiers have been developed based on the training dataset: (1) using only three categories based on an equalized training corpus ($3ceq$); (2) one which also uses the three categories but all book reviews ($3call$); (3) using the original five categories by *Amazon* with the equalized corpus ($5seq$); and (4) the unchanged training corpus with five categories and all reviews ($5sall$).

Besides the chosen approach, a number of different online or cloud deep learning methods are available. I decided for two reasons not to use any of these. First, most of these services are not free of charge, and second, the applied algorithm remains in most cases a black box. Therefore, the user is unable to interfere with or interpret how the result is produced. Google Prediction API is well known. Besides the Google service, Thomson Reuters Open Calais API,

[16] The applied code is orientated on the *SVM* tutorial from Alexandre Kowalczyk on http://www.*SVM*-tutorial.com/2014/11/*SVM*-classify-text-r/, accessed on 1 December 2016 and later adjusted step by step.
[17] Undocumented test runs for the Naive Bayes classifier have been performed. However, the overall quality of the results unsatisfying and the algorithm is therefore not presented in this study.

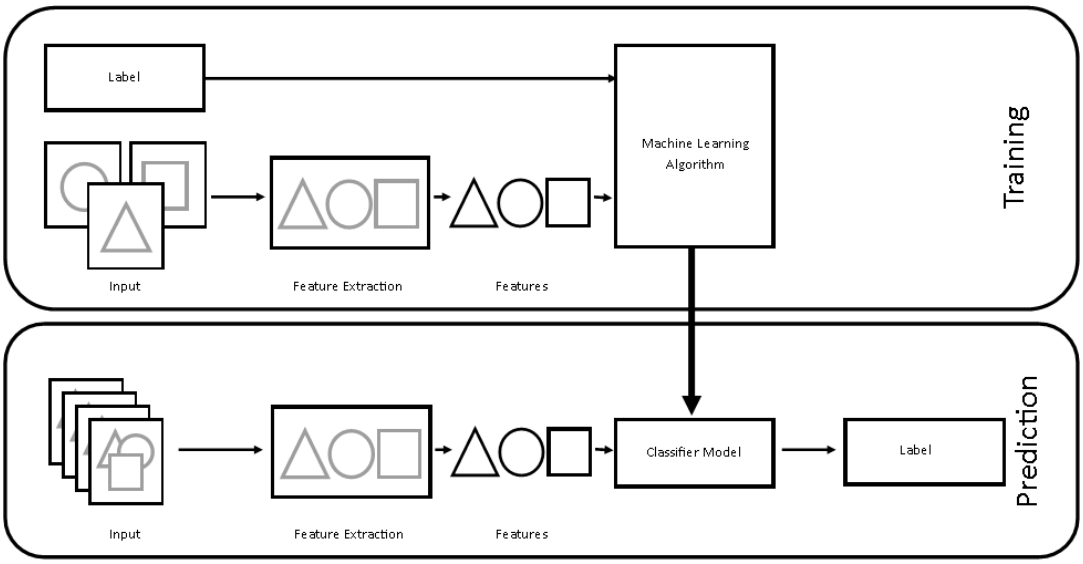
Amazon Web Services, BigML and Smart Autofill are available. I have run several trials with the last two methods since they are free of charge - in a basic version. Nonetheless, I encountered processing issues in terms of the amount of data. With Smart Autofill a maximum number of 15,000 entities can be simultaneously processed.

In the following section, I will introduce a simple probit model, where the textual sentiment indicators will be used to predict the turning points.

5.4.1 ALGORITHMS

All algorithms share in general the same structure, which consists of two steps, a training and a prediction step. In the first step, an algorithm is trained based on a set of different annotated or labelled documents. This set of documents is called the training corpus. Afterwards, the trained algorithm is applied to a new set of documents, the test corpus. This corpus enters the prediction process without any labels.

Figure 5:3 - Graphical illustration of the supervised learning approach



Note 5.7: The figure illustrates the overall process of the supervised learning approach. The approach consists of two stages a training and a prediction stage. In the training step, a number of labelled documents is used to train the machine learning algorithm. The quality of these algorithms can be checked since the corresponding label for each document is known. The trained algorithm is then tested in the second step. Here a new dataset is labelled with the help of the trained algorithms. Only if the labels for this new dataset are known, a quality check can be performed.

Figure 5:3 shows that the prediction process is more a classification issue than a labelling issue. To verify how good the developed classifier is, the produced classifications will be compared to the withheld existing labels. This, however, is only possible if the test corpus has been annotated in any other way.

Since the *Amazon* book reviews carry a corresponding label, which allows a comparison of the new labels and the old labels, I have divided the training corpus into 80% and 20%. The model is trained on the 80% of the labelled documents, and it is then tested on the remaining 20%. Using this method guarantees that performance measures can be generated. In the next step, the trained models are applied to the articles of the overall news corpus.

Note to the reader:

Please refer to section 8.1.1 in the Appendix for a more comprehensive empirical framework section. The nine different algorithms and their mathematical structures are explained here in more detail. Throughout the following chapter, I will refer to various sections in the Appendix to provide a better understanding of the methods.

5.4.2 PROBIT MODELS

Probit models are an easy way to detect changes within the underlying market. The calculation of the referring probabilities and the application of this model class has been widely used in real estate. In Tsolacos et al. (2014), a probit model is applied to a range of leading indicators and compared to the results of a Markov switching model. Similar to chapter 4.5.1, it was my intention to keep the model framework simple in order to solely focus on the leading series. I am aware, that the models lag several control variables such as the GDP, the interest rate or other real estate market factors. Focusing solely on the textual indicators their magnitude and influence on the dependent variable becomes clearer.

The dependent variable in probit models is dichotomous and takes the values 0 or 1. I have decided to use the change of the *MSCI* all property growth rate for all assets and offices $(MSCI)$. The two dependent variables are given on a monthly level from January 2004 to February 2017, with a total of 158 observations.

$$\Pr[MSCI_t = 1] = \Phi\left(\sum_i f(textSent_{t-i})\right) \qquad \textit{Equation 5:1}$$

with $MSCI_t = 1$ if the monthly overall growth rate is negative at time t and vice versa. The different textual sentiment indicators $f(textSent_{t-i})$ are applied to the model, with the later in this study to determine lag structure, via the use of the AIC.

I will not apply all constructed indicators, but those which have been proven statistically relevant. Pr states the probability forecast for the dependent variable at time t, given the cumulative density function of the normal distribution.

Equation 5:2 and Equation 5:3 state the empirical models,

$$\Pr[MSCI\_cg\_aa\_ap_t = 1] = \alpha + \sum \beta_i \, textSent_{t-i} + \varepsilon_t \qquad \textit{Equation 5:2}$$

$$\Pr[MSCI\_cg\_aa\_ao_t = 1] = \alpha + \sum_i \beta_i \, textSent_{t-i} + \varepsilon_t \qquad \textit{Equation 5:3}$$

with α and β_i being coefficients, which will be estimated. ε_t refers to the normally distributed error term. The textual sentiment represented by ($textSent_{t-i}$). The dependent variables, as dichotomous growth rates for all assets and all properties $MSCI\_cg\_aa\_ap_t$ and respectively $MSCI\_cg\_aa\_ao_t$, for all offices.

5.5 THEORETICAL EXPECTATIONS

One central question of this thesis is: What is the very nature of the underlying sentiment indicator? As discussed before the literature differentiates between direct and indirect sentiment indicators. I have further discussed an online search volume indicator, which incorporates elements of the other two classes. In this chapter, I introduce textual sentiment indicators based on news articles. Different to the previous chapter, this new set of indicators is constructed with the help of supervised learning algorithms. Given the previously presented results and the discussed shortcomings, I assume that the sentiment extracted from a large number of articles will provide sufficient information about the market sentiment.

In this study, I use multiple newspapers to avoid a biased view on market development. I assume that the reader will be influenced by the content and that he will adjust to the new situation as described in the articles by changing his behaviour.

Looking at the wording of the articles, someone would assume that when the content of the articles has a positive message, the reader would have an optimistic opinion about the discussed topic and vice versa. Unfortunately, the actual picture differs and reveals a stronger bias toward the negative information in texts. Garcia (2013) has performed an extensive study of financial news articles. The author identified that journalists tend to put more focus on adverse events. Different to Shiller (2000) who assumed that, based on behavioural finance theories, both positive and negative events should be equally present in the media, Garcia (2013) found a highly non-linear relationship between market returns and the content of news articles. Negative stock market developments are covered much more heavily and, in these phases, more extreme language is used, even when the current situation is actually not as bad as described by the journalists.

One explanation can be found in a different theory of behavioural finance, which states that it is easier to miss a gain than lose actual money. That was proven with the prospect theory by Kahneman and Tversky (1979), and it leads to the fact that a textual sentiment indicator based on news articles should be able to pick up negative events much more efficiently, but will react to positive developments not as rapidly. Furthermore, the upward movements of the textual sentiment indicator in times of positive developments will be more moderate due to the language used.

A valid question which arises from this observation is: Why? Garcia (2013) is not the first who has observed this asymmetry. Tetlock et al. (2008) stated that, when dealing with textual

analysis, negative words have a stronger impact on the sentiment and should, therefore, be used in the first place.

A rather evolutionary explanation can be found in the fact that negative events are essential for the human species and its survival. The possible danger which threatens our lives has a substantial impact on our behaviour. The human brain is trained always to scan our environment for possible threats and then adjust our behaviour in the case where it sees a reason to do so. According to Soroka and McAdams (2015), this could be the reason why people are drawn to negativity and put more emphasis on these events – they need to be informed. In an experiment, the authors have shown that, even when people report that they would prefer more positive news, they read the negative news instead. Garcia (2013) offered a different explanation and assumed journalists to be either demand or supply led.

John Authors (2017), a *Financial Times* journalist, lately commented on this observation and offered two different perspectives. He states that it would be much more devastating to encourage investors to invest money and be wrong at the end and therefore responsible for the loss of others, than convincing them not to invest. The second reason which is offered for the observed negativity bias is that Authors sees himself and his fellow journalists as at the forefront of protecting people and investors against people who want them to oversell.

5.6 RESULTS

The results section is separated into two parts. The first part will use the Amazon book reviews to train the different sentiment measures. The second part will combine the two previously used methods of word lists and supervised learning methods.

5.6.1 APPLICATION OF AMAZON BOOK REVIEWS

The following sections will discuss (1) the performance of different algorithms over the different training sets; (2) graphical analysis of the produced textual sentiment for the different classifiers and the different sub-corpora; (3) an application of the constructed sentiment indices into a probit framework. Finally, I will present (4) a series of robustness checks, which will be used to confirm my findings and underline the results.

5.6.1.1 PERFORMANCE ANALYSIS

The performance of the different algorithms is judged in two stages. The first stage relies on the split of the training data into actual training and initial training data. As previously discussed, the advantage of the training data is that all instances are labelled, and a judgement about the performance of the algorithms can be made.

The second stage of the performance analysis is based on personal judgement and personal assumptions. Since the actual test dataset (news articles) are not labelled, the output of the different algorithms cannot be judged against any pre-knowledge. To justify how good an algorithm performs, the individual results will be analysed in a graphical and statistical way.

5.6.1.1.1 TRAINING DATA: PERFORMANCE ANALYSIS

In order to estimate how well the different classifiers, perform, three measures for each of the applied algorithms were calculated: *precision, recall and the f-score*. These measures can be calculated with the withdrawn training dataset (20%). As described the algorithms are trained on 80% of the annotated dataset. Afterwards, these classifiers are applied to the withdrawn share of data to generate a label for each instance. Since the original and assumed correct label for the withdrawn dataset is known, a comparison between the classifiers result and the expected result can be made.

Precision and recall are widely used in the analysis of search quality. The question is, how good is the output regarding a particular topic within a dataset? In this case, the newly labelled records consist of 1, 0 and (-1). Each of the classes is then compared to the expected values.

Looking only at one class at a time, all records of one class in the newly labelled dataset are retrieved. These retrieved records are likely to incorporate wrongly labelled or irrelevant instances. Precision is based on the number of relevant records, or in other words these records which are true positive (TP) are divided by the total number of retrieved records, including these records which are given as belonging to a class but are false ($false\ positive\ =\ FP$).

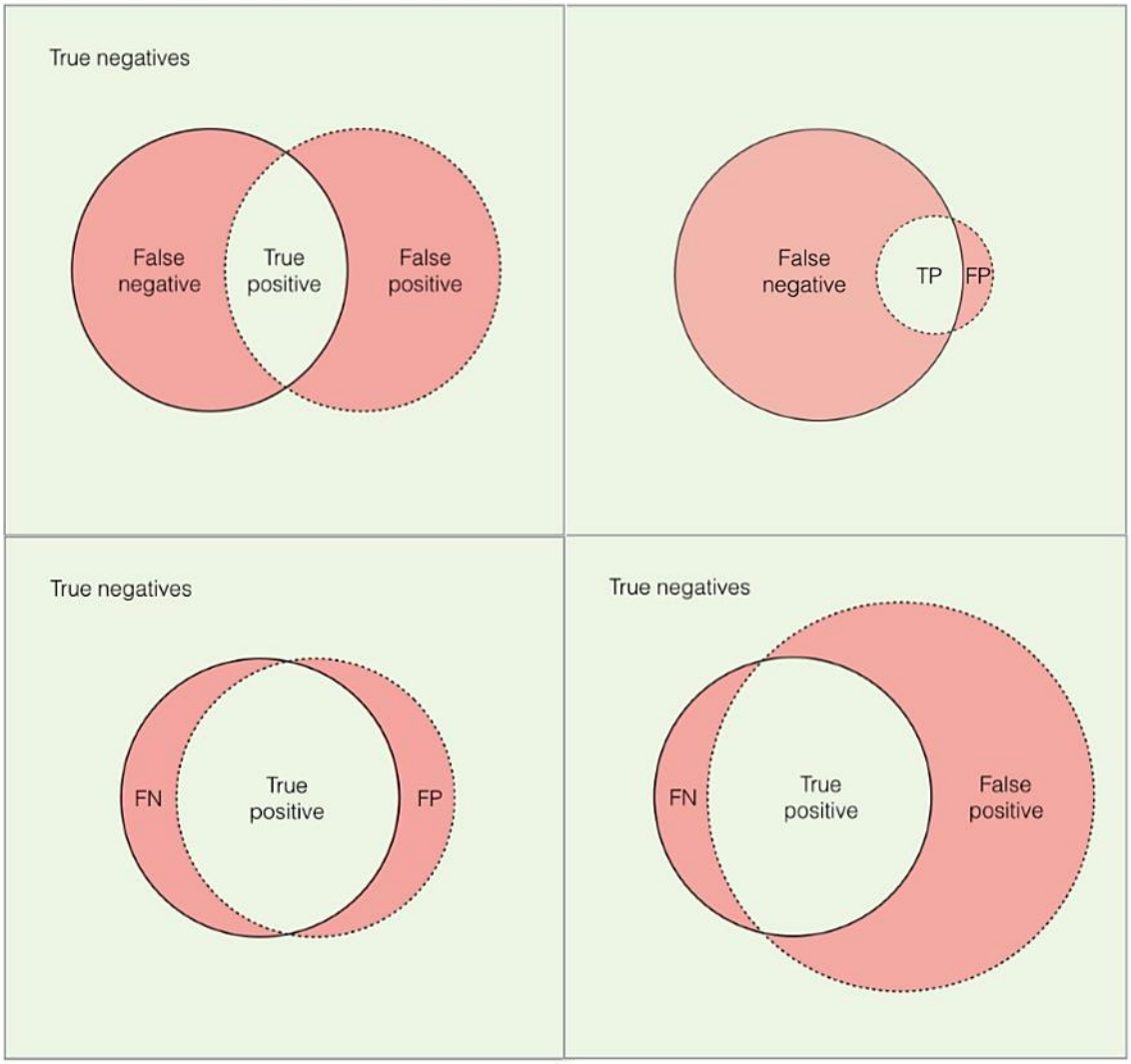
$$Precision\ (P) = \frac{TP}{(TP + FP)} \qquad \textit{Equation 5:4}$$

The second measure is recall, also known as sensitivity. Different to precision, recall states how many instances are correctly specified, based on the total number of expected instances in a class. The ratio is therefore based on the truly positive records, which were retrieved, and on those records, which should have been retrieved ($false\ negative\ =\ FN$), since they actually belong to the class of interest. In other words, recall presents the percentage of how many instances are actually correctly classified.

$$Recall\ (R) = \frac{TP}{(TP + FN)} \qquad \textit{Equation 5:5}$$

Figure 5:4 illustrate the intuition behind the two measures recall and precision. The second picture on the top on the right-hand side shows a low recall value with a high precision. Here the algorithm has identified a low number of entities (dashed line) which belong to the corresponding class, but most of them are correctly classified (more TP than FP). The fourth picture at the bottom on the right-hand side shows the other extreme. Here a good recall value has been reached with a low precision. The algorithm has identified a large number of entities, which belong to the class (TP), but also identifies many entities which do not belong to the class (FP). The picture on the left-hand side at the bottom shows the desired outcome. Here both values precision and recall are fairly high, meaning that many entities are correctly classified as belonging to the class and they are actually belonging to the class.

Figure 5:4 - Graphical illustration of precision and recall



Note 5.8: The four figures illustrate the relationship between the two measures for a single class. The dotted circle shows the results of the classifier. The full circle shows the actual instances belonging to the class. The overlap represents the correctly specified instances, the True Positives.[18]

The last measure is the $f - score$, also called the $F1 - score$. The score is a weighted average of the two previous measures and provides roughly the average between the precision and the recall. The score ranges between 1 (best) and 0 (worse).

$$F - score = \frac{2 * P * R}{P + R}, \qquad \textit{Equation 5:6}$$

[18] Pictures taken from https://medium.com/@klintcho/explaining-precision-and-recall-c770eb9c69e9, accessed on 23.11.2018.

with P for precision and R for recall. The measure allows to draw a conclusion of the tradeoff between the weight of precision and recall. This however, depends on the target the user wishes to achieve. For instance, if the initial outcome of the algorithm suggests a precision of 80% with a recall of 15%, the $F - score$ would be 25.3%. By adjusting the algorithm, I reach a slightly worse precision score of 75% but achieve an increase in recall of 5%, the harmonic measure of F increases to 31%. Therefore, the question is, if the drop in the precision value is worth it. In this case, yes, it would be worth to proceed.

Table 5:5 and Table 5:6 illustrate the performance measures for the different algorithms. Table 5:5 shows the results for the unchanged training corpus, with five classes. The first table displays the results for the whole corpus with 224,394 book reviews. The lower table presents the results for the equalized corpus over the five categories.

It can be seen that some algorithms have produced unsatisfying or even no results at all; these algorithms have been grey shaded (*SVM*[19] or *BOOSTING*[20]). It becomes clear that the algorithms perform less efficiently with multiple classes and with a large training dataset. The production of the *5sall* performance measures has taken much more computing time than all other tries. In comparison, it has also led to the most mediocre results.

None of the remaining classifiers has reached a high performance (above 0.6) for the first analysis. It can also be seen that the highest $F1 - score$ was produced by the *TREE* classifier. The reason for this is manifested in the fact that the classifier has labelled all tested entities to be class 5. This has produced higher precision and higher recall values. This unfortunate result is further confirmed by the low overall recall value, since the perfect recall value is divided by the number of classes. Only the *MAXENT*[21] classifier reaches a value above 0.3, meaning that more than one-third of the instances have been labelled correctly.

The lower part of Table 5:5 shows some improvement. None of the algorithms has a tendency towards the positive classes (4 and 5). The equal training corpus allows for a more stable distribution over the different classes, which seems to improve the classifiers. Further, the corpus is much slimmer which reduces the calculation time tremendously. All but the *TREE* classifiers have produced results over the five classes. Even though none of them has reached a higher $F1 - score$ value than 0.418 (*SVM*), the results are stable over the different classes.

[19] For further explanations regarding the SVM classifier, please refer to chapter 8.1.1.1.
[20] For further explanations regarding the BOOSTING classifier, please refer to chapter 8.1.1.6.
[21] For further explanations regarding the MAXENT classifier, please refer to chapter 8.1.1.2.

Recall values have also been improved throughout the different classifiers, with the *RANDOM FOREST*[22] classifier reaching a value of 0.418.

After I reduced the maximum number of possible categories to three, the performance over the different classifiers improved significantly (Table 5:6). In the first part of the table, where all reviews have been used for the training purpose, seven out of nine algorithms produced acceptable results. The highest overall precision (0.703) and the highest overall $F1 - score$ (0.400) was reached by the *RANDOM FOREST* classifier. *GLMENT*[23], *SLDA*[24] and *BAGGING*[25] also generated precision values above 0.5. Yet only the *MAXENT* classifier was able to allocate more than 50% of the records correctly.

This picture is further improved over the balanced training corpus. All but the *NNET* [26] and *TREE*[27] approach produced consistent results. All precision values are above 0.5, where *GLMENT* reaches a value of 0.730. Yet, I assume that the measures for *SVM*, *MAXENT* and the *RANDOM FOREST* approach are more stable, with $F1 - scores$ above 0.5.

To conclude, both *TREE* and *NNET* produced the lowest quality over the four tries, which is, with regards to the neural network approach, somewhat disappointing, since it seems to be the most promising algorithm.[28]

[22] For further explanations regarding the RANDOM FOREST classifier, please refer to chapter 8.1.1.8.
[23] For further explanations regarding the GLMENT classifier, please refer to chapter 8.1.1.4.
[24] For further explanations regarding the SLDA classifier, please refer to chapter 8.1.1.3.
[25] For further explanations regarding the BAGGING classifier, please refer to chapter 8.1.1.7.
[26] For further explanations regarding the *NNET* classifier, please refer to chapter 8.1.1.9.
[27] For further explanations regarding the *TREE* classifier, please refer to chapter 8.1.1.5.
[28] The current literature and other applications of machine learning rely heavily on the Neural Network approach. It seems promising in the sense, that complex calculations can be performed by multiple layers or neuron. For instance, Google Translate has been massively improved by a change of the underlying algorithm to NNET. (please refer to Wu et al. (2016).)

Table 5:5 - Performance analysis: five classes

Training Model	Class	SVM			MAXENTROPY			GLMENT			SLDA			BAGGING			BOOSTING			RANDOM FOREST			NNET			TREE		
		Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score
5 categories all book reviews (5s_all)	1				0.360	0.390	0.370	0.290	0.000	0.000	0.330	0.140	0.200	0.390	0.060	0.100				0.490	0.060	0.110	-	0.000	-	-	0.000	-
	2				0.160	0.140	0.150	-	0.000	-	0.190	0.070	0.100	0.090	0.000	0.000				0.280	0.010	0.020	-	0.000	-	-	0.000	-
	3				0.390	0.200	0.260	0.540	0.050	0.090	0.520	0.060	0.110	0.420	0.140	0.210				0.470	0.190	0.270	0.300	0.260	0.280	-	0.000	-
	4				0.340	0.210	0.260	0.350	0.030	0.060	0.390	0.080	0.130	0.350	0.070	0.120				0.370	0.090	0.140	-	0.000	-	-	0.000	-
	5				0.680	0.850	0.760	0.600	0.990	0.750	0.610	0.960	0.750	0.620	0.950	0.750				0.630	0.950	0.760	0.650	0.940	0.770	0.580	1.000	0.730
Overall					0.386	0.358	0.360	0.445	0.214	0.225	0.408	0.262	0.258	0.374	0.244	0.236				0.448	0.260	0.260	0.475	0.240	0.525	0.580	0.200	0.730
5 categories equal training corpus	1	0.500	0.570	0.530	0.510	0.520	0.510	0.480	0.490	0.480	0.520	0.490	0.500	0.340	0.640	0.440	0.240	0.830	0.370	0.460	0.580	0.510	0.250	0.030	0.050	0.660	0.100	0.170
	2	0.430	0.390	0.410	0.410	0.400	0.400	0.400	0.350	0.370	0.420	0.360	0.390	0.420	0.260	0.320	0.270	0.140	0.180	0.400	0.480	0.440	0.320	0.770	0.450	-	0.000	-
	3	0.370	0.260	0.310	0.380	0.220	0.280	0.270	0.230	0.250	0.300	0.280	0.290	0.400	0.090	0.150	0.490	0.080	0.140	0.390	0.200	0.260	0.250	0.010	0.020	-	0.000	-
	4	0.420	0.430	0.420	0.380	0.440	0.410	0.410	0.400	0.400	0.400	0.410	0.400	0.360	0.390	0.370	0.440	0.100	0.160	0.420	0.410	0.410	0.230	0.000	0.000	-	0.000	-
	5	0.430	0.413	0.418	0.420	0.395	0.400	0.390	0.368	0.375	0.410	0.385	0.395	0.380	0.345	0.320	0.360	0.288	0.213	0.418	0.418	0.405	0.263	0.203	0.130	0.660	0.025	0.170
Overall		0.430	0.413	0.418	0.420	0.395	0.400	0.390	0.368	0.375	0.410	0.385	0.395	0.380	0.345	0.320	0.360	0.288	0.213	0.418	0.418	0.405	0.263	0.203	0.130	0.660	0.025	0.170

Note 5.9: The table above illustrates the three performance measures for the nine different algorithms. The results are based on the original training dataset with five categories (1star – 5stars), within a total of 224,394 book reviews. The first table uses the whole training corpus (5s_all), while the second table uses the balanced training corpus (5s_eq) with 37,740 reviews. Each algorithm has been trained on 80% of these reviews, and the displayed results are generated on the remaining 20%. For each of the algorithm's precision, recall and the f-score are calculated on a class level. The "overall" row illustrates the average over the different classes. Grey shaded algorithms have not produced good results, they either failed to distribute the entities over the classes, or I was forced to cancel the prediction process.

Table 5:6 - Performance analysis: three classes

Training Model	Class	SVM			MAXENTROPY			GLMENT			SLDA			BAGGING			BOOSTING			RANDOM FOREST			NNET			TREE		
		Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score
3 categories all book reviews (3c_all)	-1				0.440	0.450	0.440	0.520	0.020	0.040	0.420	0.160	0.230	0.500	0.130	0.210	0.290	0.220	0.250	0.620	0.080	0.140	0.400	0.370	0.380	-	0.000	-
	0				0.430	0.160	0.230	0.810	0.030	0.060	0.630	0.040	0.080	0.680	0.100	0.170	0.560	0.080	0.140	0.720	0.110	0.190	-	0.000	-	-	0.000	-
	1				0.810	0.920	0.860	0.750	1.000	0.860	0.770	0.980	0.860	0.770	0.980	0.860	0.780	0.950	0.860	0.770	0.990	0.870	0.790	0.970	0.870	0.750	1.000	0.860
Overall					0.560	0.510	0.510	0.693	0.350	0.320	0.607	0.393	0.390	0.650	0.403	0.413	0.543	0.417	0.417	0.703	0.393	0.400	0.595	0.447	0.625	0.750	0.333	0.860
3 categories equal number of book reviews	-1	0.720	0.780	0.750	0.720	0.740	0.730	0.690	0.740	0.710	0.700	0.710	0.700	0.590	0.790	0.680	0.470	0.920	0.620	0.650	0.840	0.730	0.700	0.790	0.740	0.430	0.930	0.590
	0	0.520	0.090	0.150	0.400	0.170	0.240	0.930	0.040	0.080	0.660	0.040	0.080	0.540	0.060	0.110	0.650	0.040	0.080	0.590	0.090	0.160	-	0.000	-	-	0.000	-
	1	0.610	0.820	0.700	0.620	0.780	0.690	0.570	0.810	0.670	0.570	0.820	0.670	0.580	0.640	0.610	0.620	0.320	0.420	0.630	0.720	0.670	0.590	0.820	0.690	0.560	0.210	0.310
Overall		0.617	0.563	0.533	0.580	0.563	0.553	0.730	0.530	0.487	0.643	0.523	0.483	0.570	0.497	0.467	0.580	0.427	0.373	0.623	0.550	0.520	0.645	0.537	0.715	0.495	0.380	0.450

Note 5.10: The table above illustrates the three performance measures for the nine different algorithms. The results are based on the modified training dataset with three categories (positive-neutral-negative), within a total of 224,394 book reviews. The first table uses the whole training corpus (3c_all), while the second table uses the balanced training corpus (3c_eq) with 37,740 reviews. Each algorithm has been trained on 80% of these reviews, and the displayed results are generated on the remaining 20%. For each of the algorithm's precision, recall and the f-score are calculated on a class level. The "overall" row illustrates the average over the different classes. Grey shaded algorithms have not produced good results, they either failed to distribute the entities over the classes, or I was forced to cancel the prediction process.

5.6.1.2 GRAPHICAL ANALYSIS[29]

After the classifiers were trained, the actual test data with 109,103 news articles were fed to the classifiers. As stated earlier, this test dataset is unfortunately not labelled, and the output results cannot be justified in a statistical manner. However, reinforcing the central hypothesis, I believe that the classifiers trained on real-estate-related *Amazon* book reviews are good enough to generate an adequate textual sentiment index. Each output was aggregated on a quarterly level for the generation of an index. The aggregated values were finally standardized and for further analysis plotted. For this graphical analysis exercise, only those algorithms are used which were able to produce unbiased results in the previous section 5.6.1.1.1. Has an algorithm classified all entities into one or less than possible categories, the algorithm, has been excluded from the following analysis. Algorithms, which have failed this initial classification process, have been highlighted in Table 5:5 and Table 5:6.

I have generated one output for each classifier based on the full article text.[30]

For comparison reasons, I have analysed the test dataset with the classical lexical approach. I used topic modelling, from the topic modelling r-package by Feinerer and Hornik (2008) and the *AFINN*, *BING* and *NRC* approaches, which are covered in the '*syuzhet*' package by Jockers (2016). *NRC* and *TM* have produced satisfying results in chapter 4.5. These indices have been also aggregated on a quarterly level and finally standardized.

The created textual sentiment indices are further separated over the five different sub-corpora as described in section 0 (all, no housing, London, 100,000 and FT).

[29] The graphical analysis is performed on a quarterly level, while the later probit analysis is performed on a monthly level.
[30] Unreported results for the analysis of the titles of each article have not produced sufficient results. My initial assumption, that the titles and the book reviews share a similar structure, was not confirmed. The classifiers rather rely on the word structure of the whole text and assign the classes based on the word frequency, therefore more words generate a more stable output.

5.6.1.2.1 ALL ARTICLES

LEXICON APPROACH

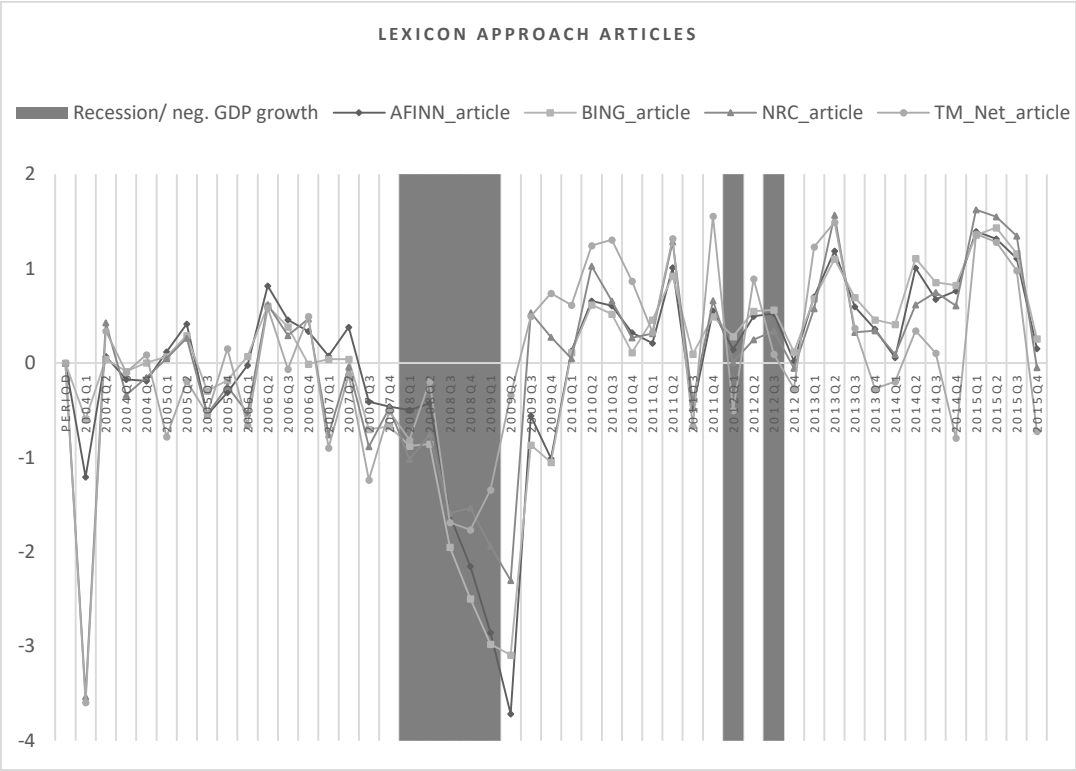
Figure 5:5 illustrates the lexicon approach for all articles. The grey shaded areas in the diagram illustrate the recession period between 2008q1 to 2009q2, as well as two quarters with negative GDP growth in the U.K. in 2012q1 and 2012q3.[31]

I assume especially over the recession period that the newspapers would have a negative coverage of the events. This should be reflected in a negative development of the textual sentiment indices.

As illustrated in the graph the four lexicon-based indicators, especially the *AFINN* indicator, show the course of the financial crisis to be a rather extreme negative development. Toward the other two periods with negative growth the indicators also have a negative development, yet, they miss the negative period of 2012q3 by one period.

[31] Data from the Office for National Statistics, https://www.ons.gov.uk/economy/grossdomesticproductgdp/timeseries/ihyq/qna, accessed on 14 December 2016.

Figure 5:5 - Lexicon approach (all articles)



Note 5.11: The figure illustrates the development of the four different lexicon-based sentiment indicators on a quarterly base. The four algorithms mirror the sentiment for the full corpus.

Table 5:7 reports the correlation results for the four lexical indicators. Most of the correlation coefficients are strongly positive, which underlines the graphical results.

Table 5:7 - Correlation analysis - lexicon approach - (all articles)

	*AFINN*_article	*BING*_article	*NRC*_article	*TM*_Net_article
*AFINN*_article	1			
*BING*_article	0.971	1		
*NRC*_article	0.846	0.802	1	
*TM*_Net_article	0.614	0.576	0.86	1

Note 5.12: The table illustrates the correlation between the different sentiment indicators constructed by the lexicon approach.

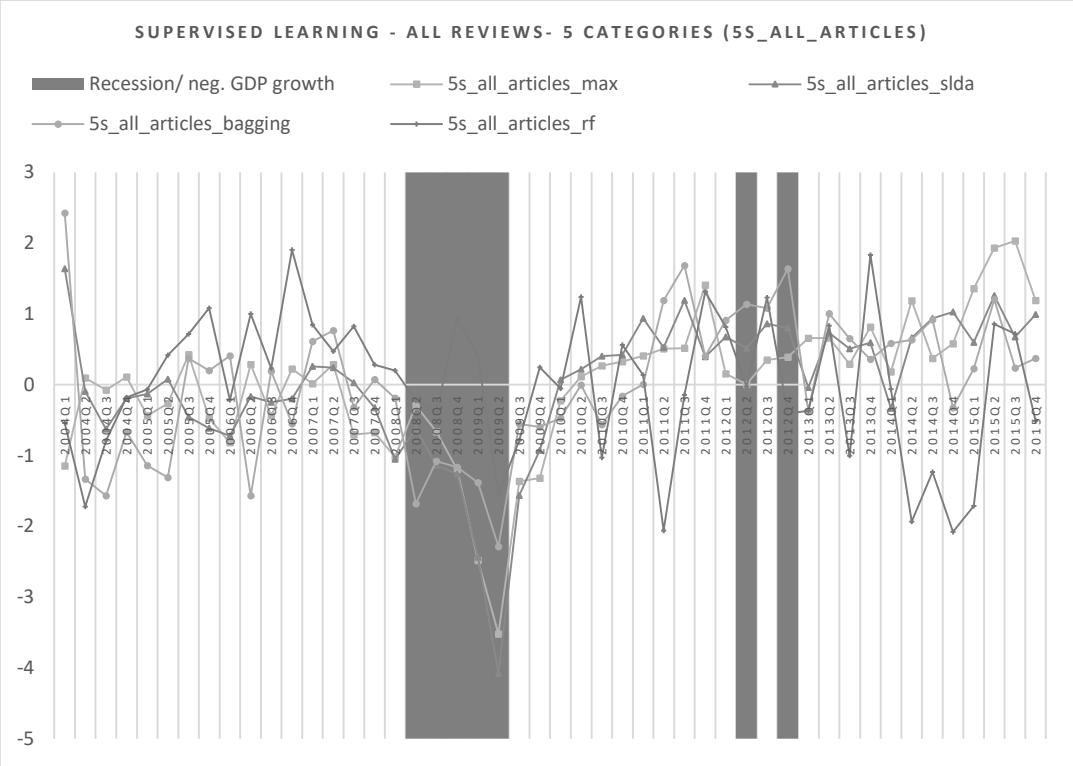
In general, the sentiment indicators based on the articles show a downward sloping trend almost two years before the recession started, which could be seen as an indicator that the wording in the articles has picked up the negative market sentiment.

It is also convincing that nearly all indicators reach their lowest level within a range of one-quarter before or after the end of the recession. This seems logical since at the end the first signs of recovery should have been present in the market and the last quarter might have been dominated by summaries of past negative events.

SUPERVISED LEARNING APPROACH

Figure 5:6 shows the results for the supervised learning algorithms, which have been trained with all reviews. The applied classifiers try to label the articles into one of five categories.

Figure 5:6 - Classifiers trained on all book reviews: five classes (all articles)



Note 5.13: The figure illustrates the development of the four supervised learning indicators, which have been trained by the full training corpus with five categories. As a test dataset, the full document corpus has been used.

The graphical results are similar to the presented results of the lexicon approach. Only the *RANDOM FOREST* ($5s\_all\_articles\_rf$) approach seems in some of the cases to be out of order. For instances during the financial crisis, the indicator produces a positive sentiment and later,

while the remaining indices predict a positive trend, *RANDOM FOREST* has a negative outlier (2011q2; 2014q2 - 2015q1).

While the correlation coefficients for the first three indicators are strongly positive, the *RANDOM FOREST* indicator shows virtually no correlation to the other three (Table 5:8).

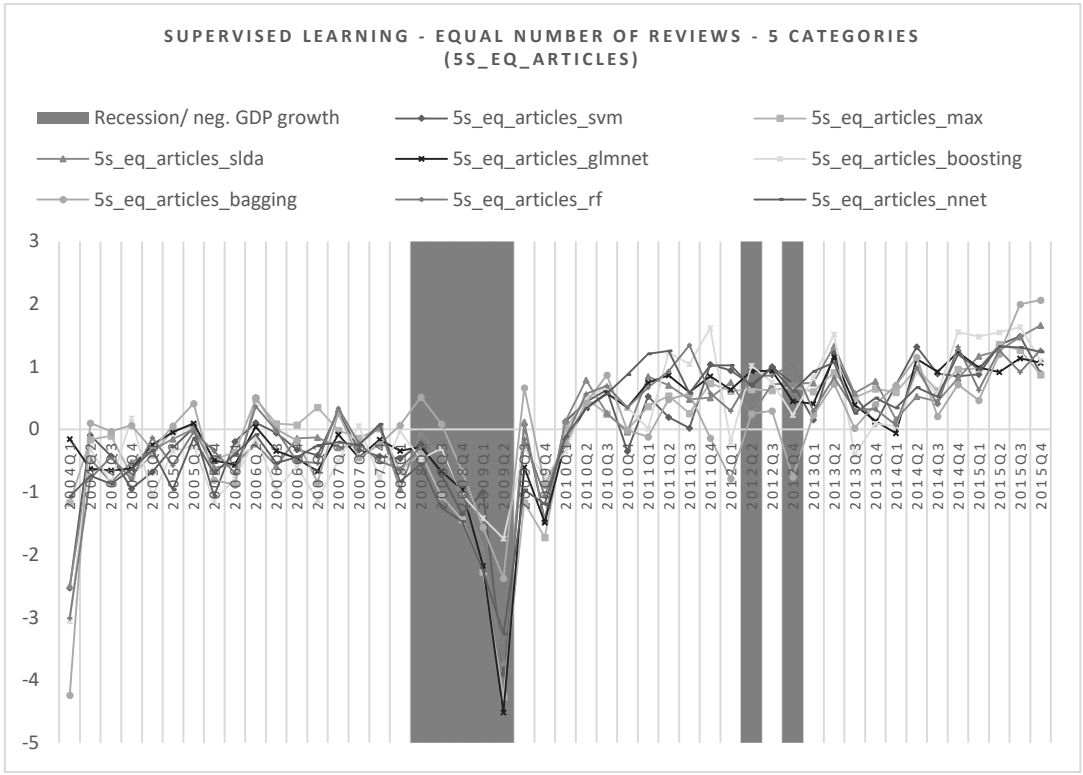
Table 5:8 - Correlation analysis - supervised learning approach - (all articles) - 5 categories - all reviews

	MAXENT (5s_all_articles_max)	SLDA (5s_all_articles_SLDA)	BAGGING (5s_all_articles_BAGGING)	*RANDOM FOREST* (5s_all_articles_rf)
MAXENT (5s_all_articles_max)	1			
SLDA (5s_all_articles_SLDA)	0.803	1		
BAGGING (5s_all_articles_BAGGING)	0.465	0.711	1	
RANDOM FOREST (5s_all_articles_rf)	0.075	0.035	0.061	1

Note 5.14:The table illustrates the correlation between the four textual indicators based on all reviews: five categories.

Figure 5:7 illustrates the textual sentiment indicators based on the equalized training corpus with five categories. The previously present tendency towards the right classes in the training data set has been removed. Due to the equalization of the five shares in the training dataset, an improvement in the results as well as in the total number of classifiers can be observed.

Figure 5:7 - Classifiers trained on an equal number of book reviews: five classes (all articles)



Note 5.15: The figure illustrates the development of the eight supervised learning indicators, which have been trained by an equalized training corpus with five categories. As a test dataset, the full document corpus has been used.

The analysis of the articles shows satisfying results. Most of the indices were able to show the expected adverse development over the course of the recession period. For the two negative quarters towards the end of my analysis period, the results are also encouraging. However, similar to the lexicon approach, those events are missed by one-quarter by most of the indicators.

Table 5:9 - Correlation analysis - supervised learning approach - (all articles) - 5 categories - equal number of reviews

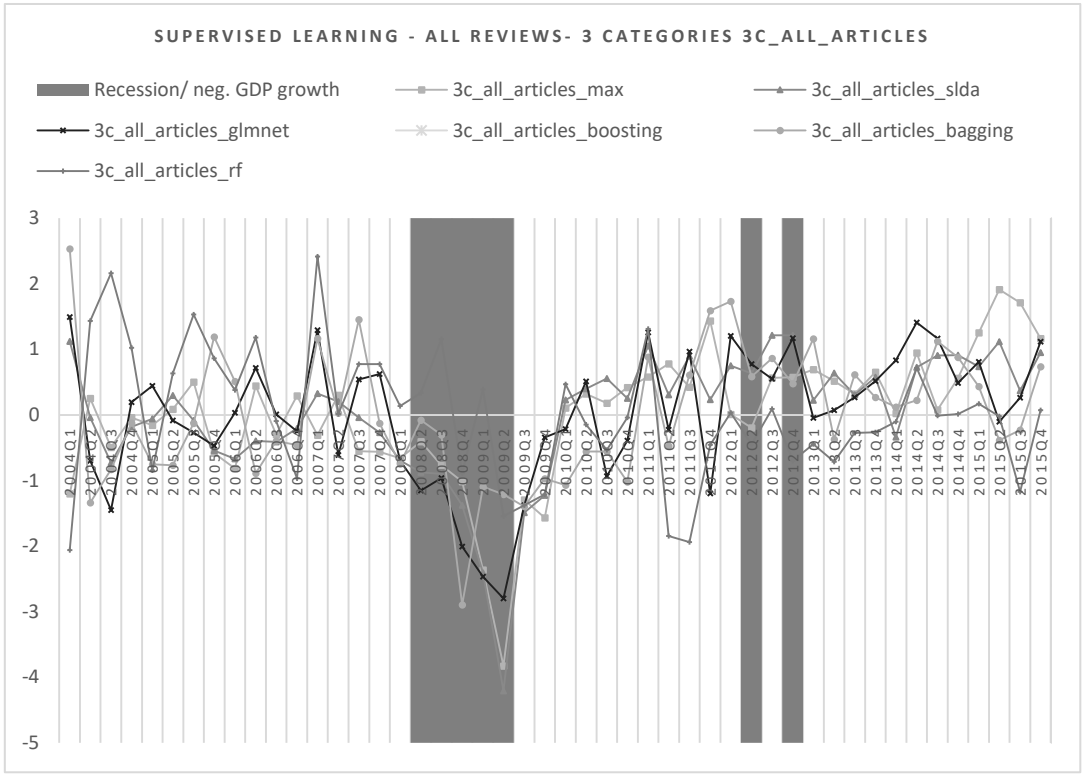
	SVM (5s_eq_articles_*SVM*)	*MAXENT* (5s_eq_articles_max)	SLDA (5s_eq_articles_SLDA)	*GLMENT* (5s_eq_articles_*GLMENT*)	BOSSTING (5s_eq_articles_*BOOSTING*)	BAGGING (5s_eq_articles_BAGGING)	*RANDOM FOREST* (5s_eq_articles_rf)	Neural Net (5s_eq_articles_*NNET*)
SVM (5s_eq_articles_*SVM*)	1							
MAXENT (5s_eq_articles_max)	0.904	1						
SLDA (5s_eq_articles_SLDA)	0.921	0.906	1					
GLMENT (5s_eq_articles_*GLMENT*)	0.867	0.93	0.879	1				
BOSSTING (5s_eq_articles_*BOOSTING*)	0.799	0.708	0.815	0.727	1			
BAGGING (5s_eq_articles_BAGGING)	0.765	0.653	0.782	0.599	0.788	1		
RANDOM FOREST (5s_eq_articles_rf)	0.908	0.87	0.925	0.864	0.84	0.807	1	
Neural Net (5s_eq_articles_*NNET*)	0.857	0.915	0.908	0.925	0.797	0.611	0.879	1

Note 5.16: The table illustrates the correlation among the eight supervised learning indicators based on an equalized training corpus: five categories.

Surprising is the initial stage of all indicators. Some show a positive development within the first quarter with a massive correction in the second, and others show a minor negative development over the same period. Until the crisis period, all indicators ranged between 1 and –1; during and after the crisis this development changed to more extreme values. The correlation analysis (Table 5:9) reveals that all indicators share a moderate to high positive correlation.

The following section will present the results for those indicators trained on only three categories. Figure 5:8 shows the outputs for the classifiers based on the full training corpus.

Figure 5:8 - Classifiers trained on all book reviews: three classes (all articles)



Note 5.17: The figure illustrates the development of the six supervised learning indicators, which have been trained by the full training corpus with three categories. As a test dataset, the full document corpus has been used.

The results are relatively acceptable compared to the other two categories. It seems that based on the graphical observation the indicators are not as much in line as for the previous equalized training corpus. During the recession period, for instance, some indicators reach their minimum up to two quarters ahead of the end of the recession, such as the BAGGING or the Random Forrest indicator.

The correlation analysis also shows that the indicators are less positively correlated as before (Table 5:10).

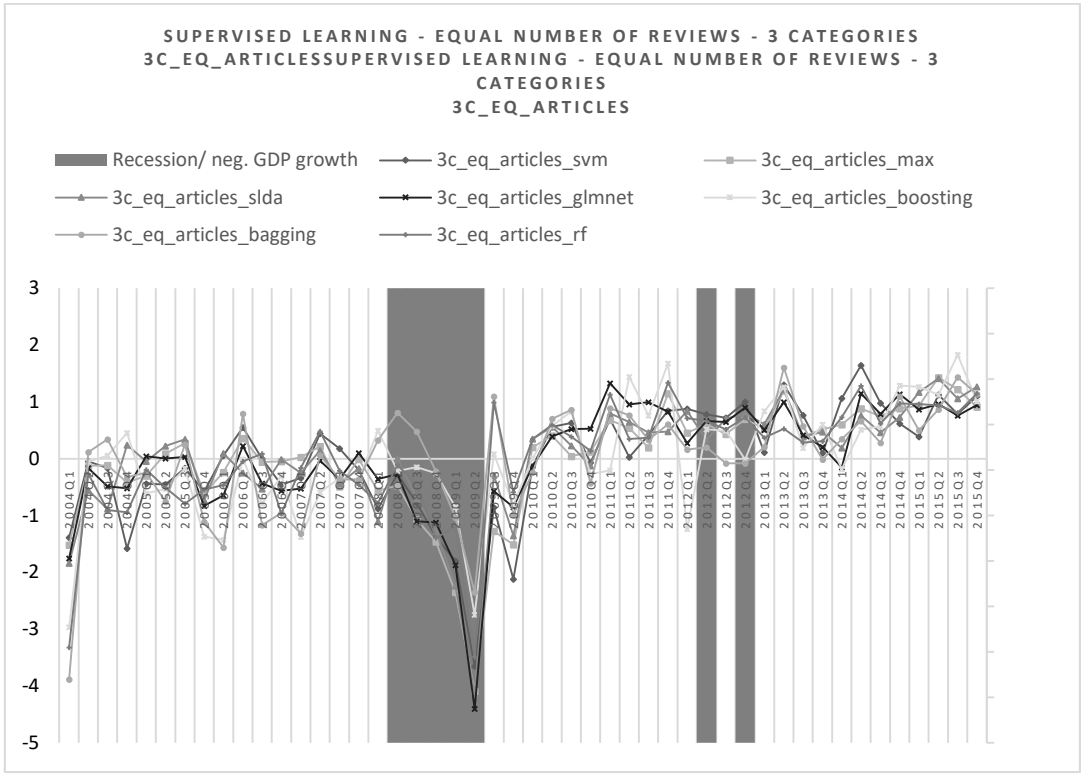
Table 5:10 - Correlation analysis - supervised learning approach - (all articles) - 3 categories - all reviews

	MAXENT (3c_all_articles_max)	SLDA (3c_all_articles_SLDA)	*GLMENT* (3c_all_articles_*GLMENT*)	BAGGING (3c_all_articles_BAGGING)	*RANDOM FOREST* (3c_all_articles_rf)
MAXENT (3c_all_articles_max)	1				
SLDA (3c_all_articles_SLDA)	0.822	1			
GLMENT (3c_all_articles_*GLMENT*)	0.531	0.793	1		
BAGGING (3c_all_articles_BAGGING)	0.282	0.552	0.632	1	
RANDOM FOREST (3c_all_articles_rf)	0.095	0.036	0.07	0.036	1

Note 5.18: The table illustrates the correlation among the five supervised learning indicators based on all reviews: three categories.

For the classifiers based on the equalized training corpus, the picture is again much more in line. All indicators start with the same positive development over the course of the first two quarters. During the recession period, all indicators show their most negative value at the end of the recession and have a sharp positive increase in 2009q3. From there onwards, the development has a positive trend with a minor dip for the two quarters with a negative GDP growth (Figure 5:9).

Figure 5:9 - Classifiers trained on an equal number of book reviews: three classes (all articles)



Note 5.19: The figure illustrates the development of the four supervised learning indicators, which have been trained by an equalized training corpus with three categories. As a test dataset, the full document corpus has been used.

This result is confirmed by strong positive correlation among the different classifiers. Only some show a moderate correlation (Table 5:11).

Table 5:11 - Correlation analysis - supervised learning approach - (all articles) - 3 categories - equal number of reviews

	SVM (3c_eq_articles_*SVM*)	*MAXENT* (3c_eq_articles_max)	SLDA (3c_eq_articles_SLDA)	*GLMENT* (3c_eq_articles_*GLMENT*)	*BOOSTING* (3c_eq_articles_*BOOSTING*)	BAGGING (3c_eq_articles_BAGGING)	*RANDOM FOREST* (3c_eq_articles_rf)
SVM (3c_eq_articles_*SVM*)	1.000						
MAXENT (3c_eq_articles_max)	0.921	1.000					
SLDA (3c_eq_articles_SLDA)	0.866	0.942	1.000				
GLMENT (3c_eq_articles_*GLMENT*)	0.881	0.935	0.927	1.000			
BOOSTING (3c_eq_articles_*BOOSTING*)	0.611	0.712	0.715	0.766	1.000		
BAGGING (3c_eq_articles_BAGGING)	0.642	0.636	0.677	0.713	0.85	1.000	
RANDOM FOREST (3c_eq_articles_rf)	0.836	0.817	0.849	0.866	0.731	0.781	1.000

Note 5.20: The table illustrates the correlation among the eight supervised learning indicators based on an equalized training corpus: three categories.

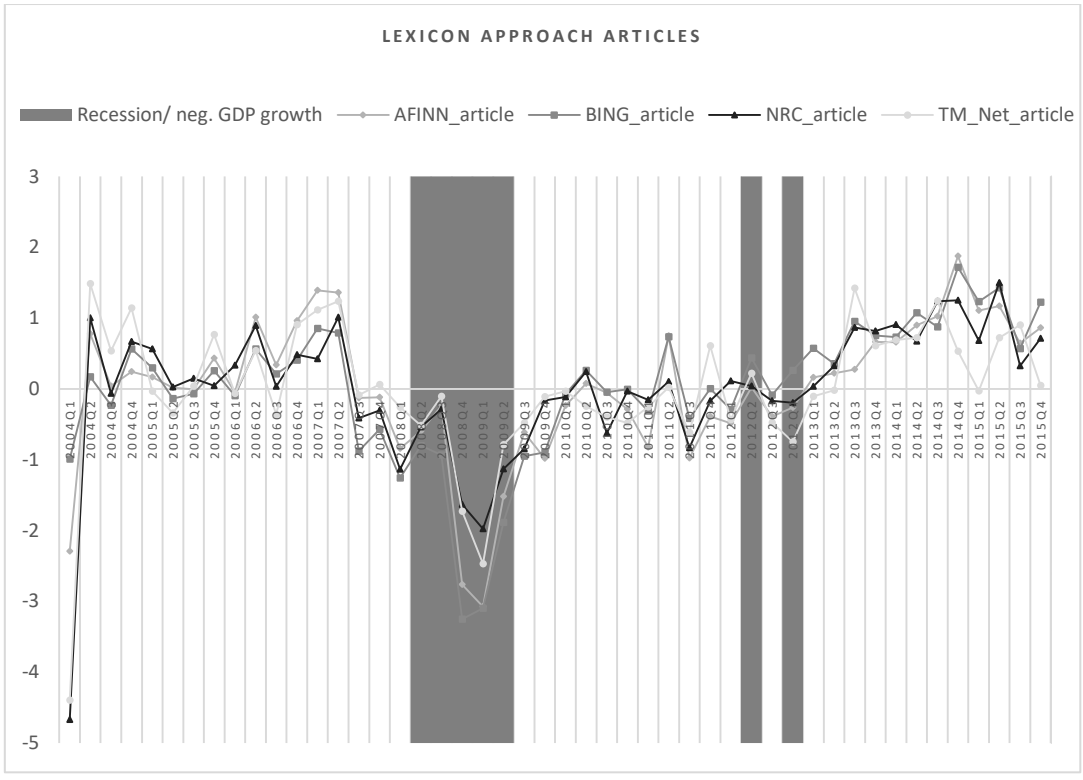
5.6.1.2.2 NO HOUSING ARTICLES

In the following analysis, housing-related worded articles have been removed from the corpus, and a textual sentiment indicator with the reduced number of articles has been produced. It was my aim to generate more commercial real estate related indicators.

LEXICON APPROACH

Starting again with the simple lexical approach (Figure 5:10), it can be seen that all four indices are in line with each other and that they pick up the recession period. However, the leading series react one to two quarters before the actual end of the recession and increase. The *TM* and the *NRC* series do miss the expected negative development at the end of the observation period.

Figure 5:10 - Lexicon approach (no housing)



Note 5.21: The figure illustrates the development of the four different lexicon-based sentiment indicators on a quarterly base. The four algorithms mirror the sentiment for the no housing sub-corpus.

The correlation analysis (Table 5:12) for these indicators reveals as expected a moderate to high positive correlation. It seems that especially the *BING* and the *AFINN* indicators share a common trend.

Table 5:12 - Correlation analysis - lexicon approach - (no housing)

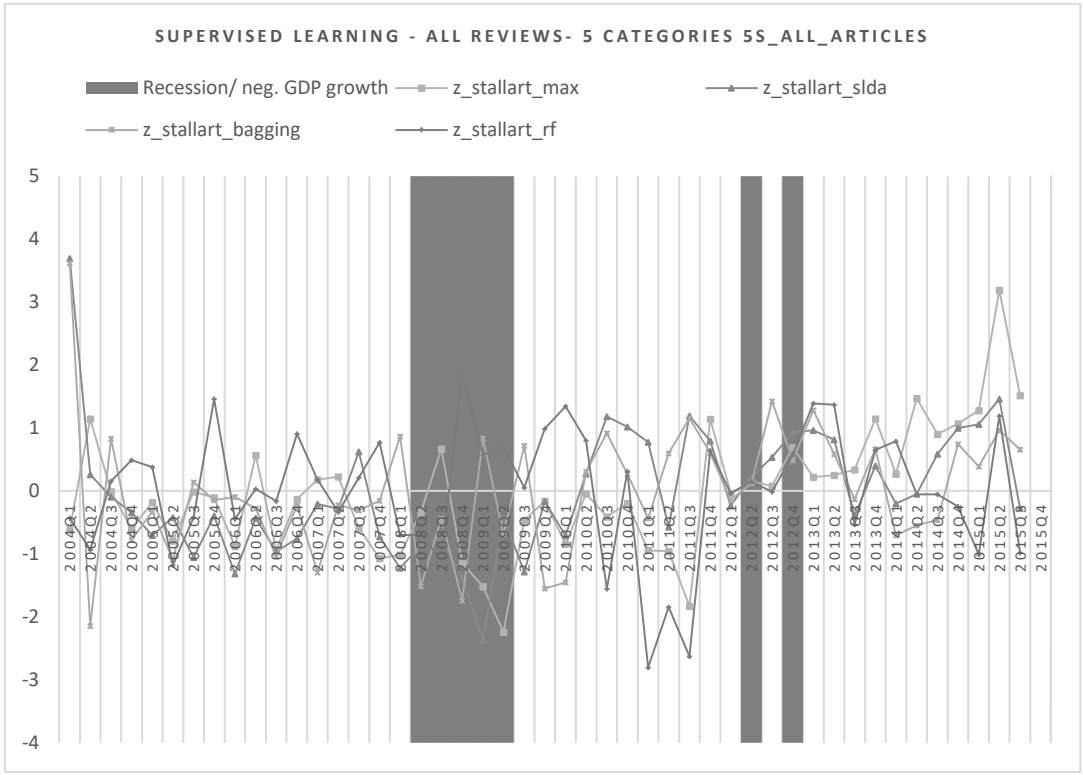
	*AFINN*_article	*BING*_article	*NRC*_article	*TM*_Net_article
*AFINN*_article	1.000			
*BING*_article	0.917	1.000		
*NRC*_article	0.846	0.747	1.000	
*TM*_Net_article	0.805	0.660	0.899	1.000

Note 5.22: The table illustrates the correlation between the four supervised learning indicators.

SUPERVISED LEARNING APPROACH

Using all the remaining articles for the five different classes, the output of the supervised learning algorithms has nothing in common with the previous analysis. The graphical illustration (Figure 5:11) shows that the indices are not in line and only some of them are able to follow the negative development over the recession period.

Figure 5:11 - Classifiers trained on all book reviews: five classes (no housing)



Note 5.23: The figure illustrates the development of the four supervised learning indicators, which have been trained by the full training corpus with five categories. As a test dataset, the sub-corpus without housing related terms has been used.

This somewhat chaotic picture of the different indicators is further confirmed in the low to moderate correlations among them (Table 5:13).

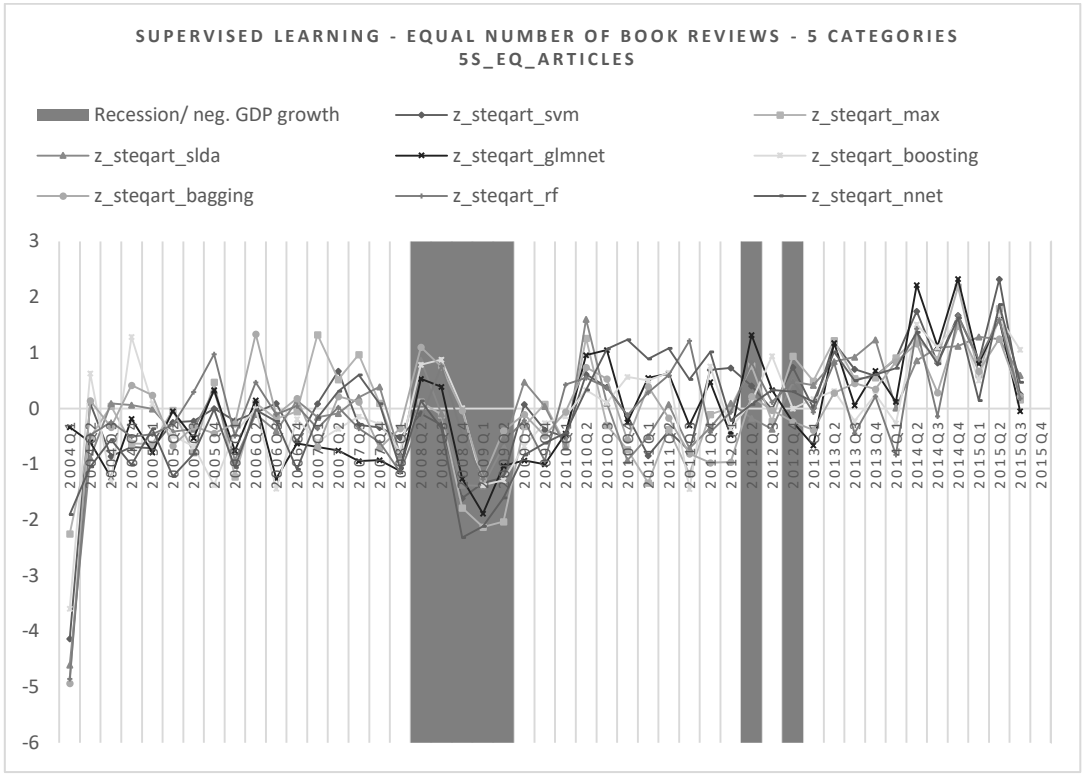
Table 5:13 - Correlation analysis - supervised learning approach - (no housing) - 5 categories - all reviews

	MAXENT (5s_all_articles_max)	SLDA (5s_all_articles_SLDA)	BAGGING (5s_all_articles_BAGGING)	*RANDOM FOREST* (5s_all_articles_rf)
MAXENT (5s_all_articles_max)	1.000			
SLDA (5s_all_articles_SLDA)	0.363	1.000		
BAGGING (5s_all_articles_BAGGING)	0.078	0.531	1.000	
RANDOM FOREST (5s_all_articles_rf)	0.143	-0.133	-0.125	1.000

Note 5.24: The table illustrates the correlation between the four supervised learning indicators based on all book reviews with five categories.

This picture improves when the balanced training corpus is applied (Figure 5:12). Here again, the indices share a common trend and also pick up the recession period. Unfortunately, they fail to show negative development over the two quarters towards the end of the selected period.

Figure 5:12 - Classifiers trained on an equal number of book reviews: five classes (no housing)



Note 5.25: The figure illustrates the development of the eight supervised learning indicators, which have been trained by an equalized training corpus with five categories. As a test dataset, the no-housing subcorpus has been used.

Table 5:14 presents the correlation analysis among the indicators. It can be observed that the correlation coefficients are now moderately or strongly positively correlated.

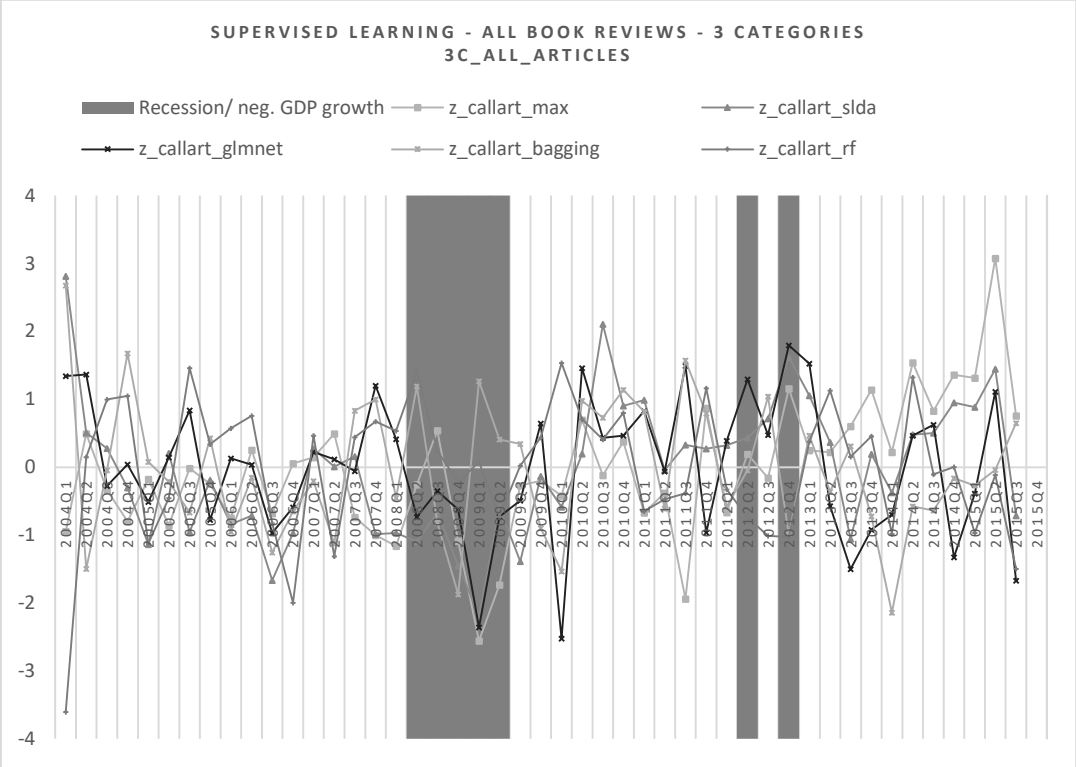
Table 5:14 - Correlation analysis - supervised learning approach - (no housing) - 5 categories - equal number of reviews

	SVM (5s_eq_articles_*SVM*)	*MAXENT* (5s_eq_articles_max)	SLDA (5s_eq_articles_SLDA)	*GLMENT* (5s_eq_articles_*GLMENT*)	BOSSTING (5s_eq_articles_*BOOSTING*)	BAGGING (5s_eq_articles_BAGGING)	*RANDOM FOREST* (5s_eq_articles_rf)	Neural Net (5s_eq_articles_*NNET*)
SVM (5s_eq_articles_*SVM*)	1.000							
MAXENT (5s_eq_articles_max)	0.828	1.000						
SLDA (5s_eq_articles_SLDA)	0.826	0.780	1.000					
GLMENT (5s_eq_articles_*GLMENT*)	0.609	0.613	0.519	1.000				
BOSSTING (5s_eq_articles_*BOOSTING*)	0.688	0.544	0.673	0.653	1.000			
BAGGING (5s_eq_articles_BAGGING)	0.728	0.580	0.800	0.522	0.773	1.000		
RANDOM FOREST (5s_eq_articles_rf)	0.753	0.582	0.751	0.578	0.645	0.785	1.000	
Neural Net (5s_eq_articles_*NNET*)	0.733	0.702	0.627	0.748	0.623	0.532	0.626	1.000

Note 5.26: The table illustrates the correlation among the eight supervised learning indicators based on an equalized training corpus: five categories.

Changing the number of classes has not produced a different result to that shown in Figure 5:11. The classification into three classes with all remaining articles has also produced a slightly chaotic picture. Yet, Figure 5:13 shows that more indicators are able to show an adverse development over the recession period. On the other hand, the starting directions, as well as the final quarters, differ among the indices.

Figure 5:13 - Classifiers trained on all book reviews: three classes (no housing)



Note 5.27: The figure illustrates the development of the five supervised learning indicators, which have been trained by the full training corpus with three categories. As a test dataset, the sub-corpus without housing related terms has been used.

The correlations among the indicators remain low to moderate and even negative in some cases (Table 5:15).

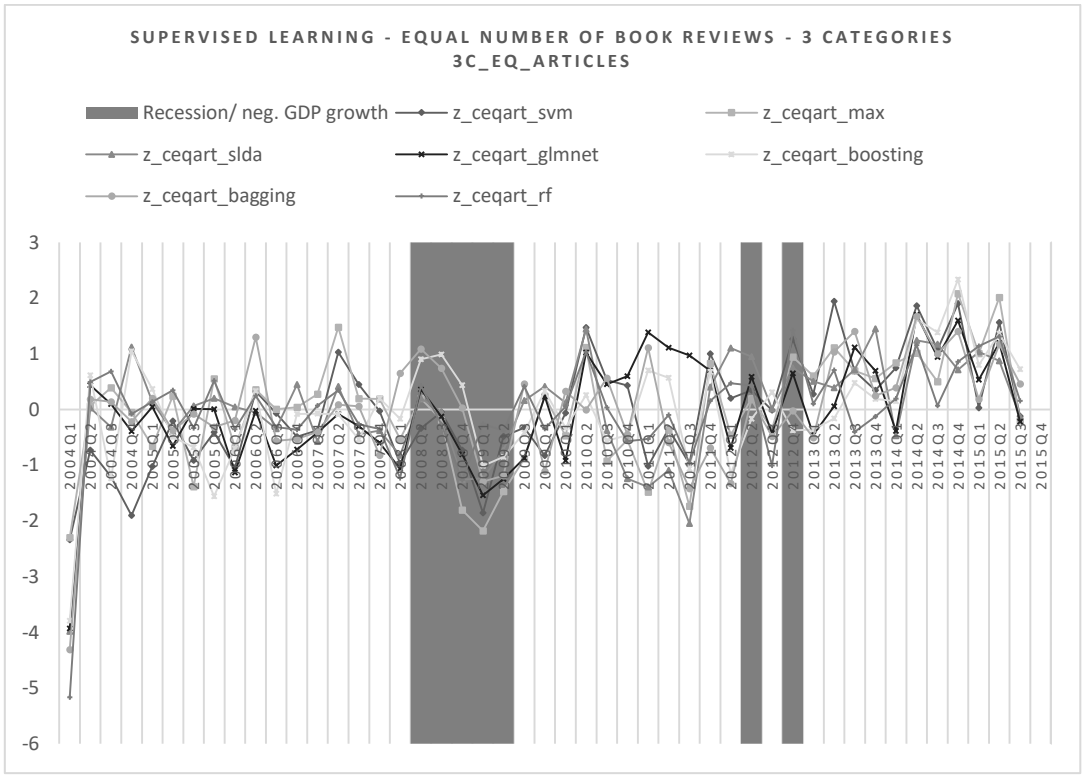
Table 5:15 - Correlation analysis - supervised learning approach - (no housing) - 3 categories - all reviews

	MAXENT (3c_all_articles_max)	SLDA (3c_all_articles_SLDA)	*GLMENT* (3c_all_articles_*GLMENT*)	BAGGING (3c_all_articles_BAGGING)	*RANDOM FOREST* (3c_all_articles_rf)
MAXENT (3c_all_articles_max)	1.000				
SLDA (3c_all_articles_SLDA)	0.425	1.000			
GLMENT (3c_all_articles_*GLMENT*)	0.131	0.556	1.000		
BAGGING (3c_all_articles_BAGGING)	-0.216	0.274	0.166	1.000	
RANDOM FOREST (3c_all_articles_rf)	0.099	-0.214	-0.110	-0.060	1.000

Note 5.28: The table illustrates the correlation between the five supervised learning indicators based on all book reviews: three categories.

Some improvement has been reached by the last analysis with the output for the three classes and the classifiers based on the equal training corpus. All sentiment indicators are in line with each other and show a similar development for both the end and the beginning of the testing period. Even though they pick up the recession period, the negative development ends up to three quarters before the actual recession ends (Figure 5:14).

Figure 5:14 - Classifiers trained on an equal number of book reviews: three classes (no housing)



Note 5.29: The figure illustrates the development of the seven supervised learning indicators, which have been trained by an equalized training corpus with three categories. As a test dataset, the no-housing subcorpus has been used.

Table 5:16 once more illustrates the correlation coefficients for the different textual sentiment indicators for the no housing subcorpus, for those classifiers which are trained on an equal number of book reviews with three classes. The correlations range between moderate to strong, showing that the indicators pick up a common trend.

Table 5:16 - Correlation analysis - supervised learning approach - (no housing) - 3 categories - equal number of reviews

	SVM (3c_eq_articles_*SVM*)	*MAXENT* (3c_eq_articles_max)	SLDA (3c_eq_articles_SLDA)	*GLMENT* (3c_eq_articles_*GLMENT*)	*BOOSTING* (3c_eq_articles_*BOOSTING*)	BAGGING (3c_eq_articles_BAGGING)	*RANDOM FOREST* (3c_eq_articles_rf)
SVM (3c_eq_articles_*SVM*)	1.000						
MAXENT (3c_eq_articles_max)	0.770	1.000					
SLDA (3c_eq_articles_SLDA)	0.594	0.717	1.000				
GLMENT (3c_eq_articles_*GLMENT*)	0.613	0.559	0.554	1.000			
BOOSTING (3c_eq_articles_*BOOSTING*)	0.501	0.543	0.591	0.718	1.000		
BAGGING (3c_eq_articles_B AGGING)	0.567	0.534	0.586	0.677	0.748	1.000	
RANDOM FOREST (3c_eq_articles_rf)	0.614	0.677	0.784	0.682	0.573	0.667	1.000

Note 5.30: The table illustrates the correlation among the seven supervised learning indicators based on an equalized training corpus: three categories.

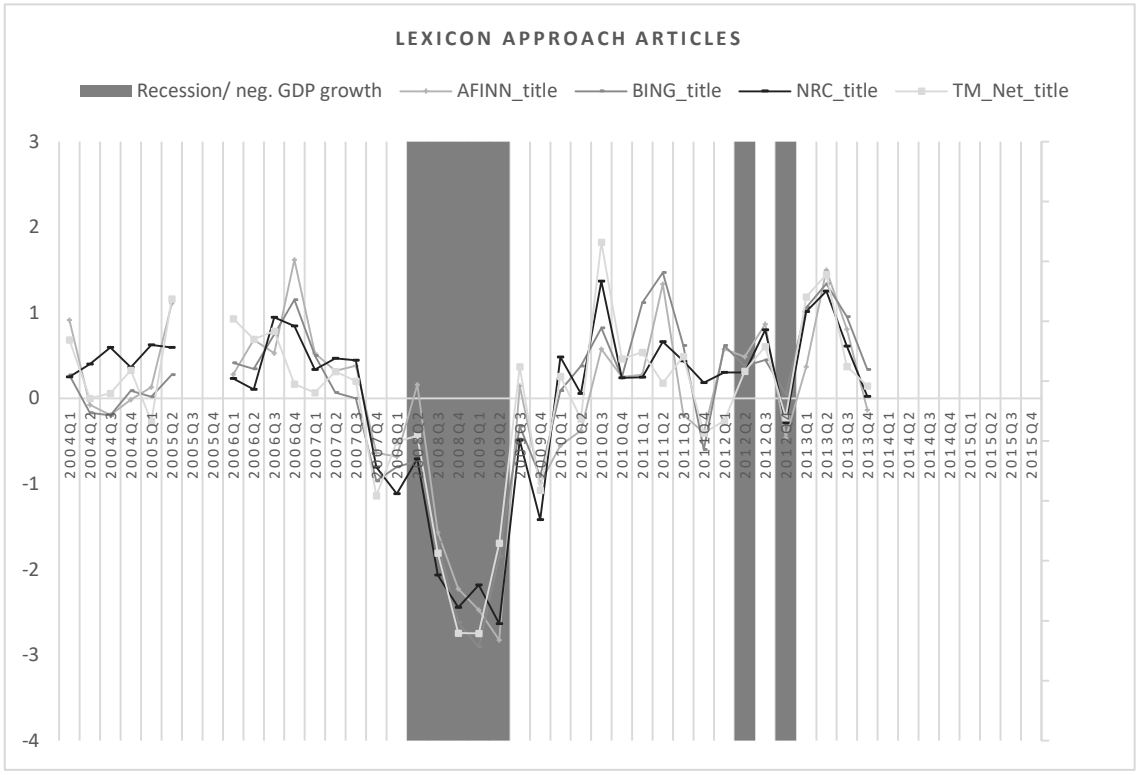
5.6.1.2.3 LONDON

The sub-corpus for London shows missing observations for two quarters in 2005 and after the fourth quarter of 2013. This observation is somewhat surprising. However, I have double checked the selected articles for the sub-corpus and have reached the same result. Besides this minor drawback, the results for the London corpus seem to be the most promising so far.

LEXICON-BASED APPROACH

Starting again with the lexical approach (Figure 5:15), it can be seen that the results do not differ from the previous ones. The indicators are able to follow the negative recession period within a range of two to one quarter, and they also pick up the negativity in the last negative quarter.

Figure 5:15 - Lexicon approach (London)



Note 5.31: The figure illustrates the development of the four different lexicon-based sentiment indicators on a quarterly base. The four algorithms mirror the sentiment for the London specific sub-corpus.

It is not surprising that the correlation among these indicators remains positive and high.

Table 5:17 - Correlation analysis - lexicon approach - (London)

	*AFINN*_article	*BING*_article	*NRC*_article	*TM*_Net_article
*AFINN*_article	1.000			
*BING*_article	0.967	1.000		
*NRC*_article	0.856	0.813	1.000	
*TM*_Net_article	0.704	0.672	0.916	1.000

Note 5.32: The table illustrates the correlation between the different sentiment indicators constructed by the lexicon approach.

Similar to the previous example (no housing related articles), those classifiers, which are trained on the biased all review training dataset with five classes, show a diversified picture (Figure 5:16). Even though the indicators follow the suggested trend in the crisis, their beginning and development until 2005q2 are out of line. The *RANDOM FOREST* index especially seems to be more extreme and in some instances behind the other indices.

Figure 5:16 - Classifiers trained on all book reviews: five classes (London)

Note 5.33: The figure illustrates the development of the four supervised learning indicators, which have been trained by the full training corpus with five categories. As a test dataset, the London specific sub-corpus has been used.

The correlation table illustrates once more that the indicators only have a weak to moderate correlation (Table 5:18).

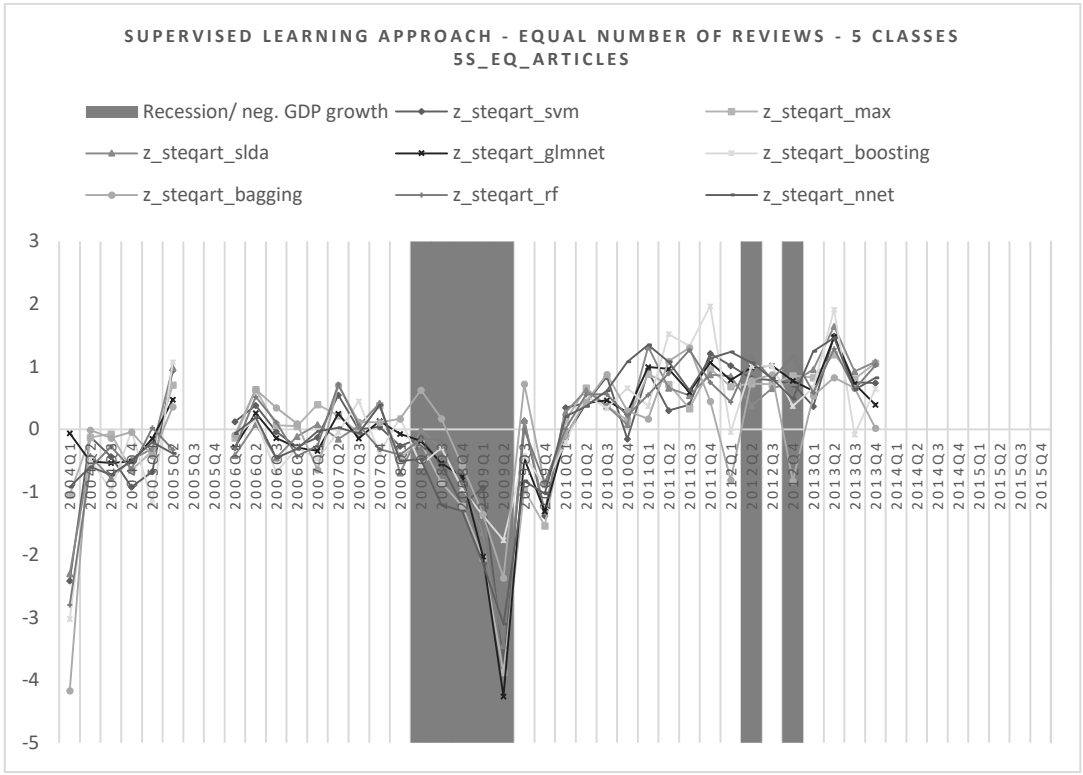
Table 5:18 - Correlation analysis - supervised learning approach - (London) - 5 categories - all reviews

	MAXENT (5s_all_articles_max)	SLDA (5s_all_articles_SLDA)	BAGGING (5s_all_articles_BAGGING)	*RANDOM FOREST* (5s_all_articles_rf)
MAXENT (5s_all_articles_max)	1.000			
SLDA (5s_all_articles_SLDA)	0.774	1.000		
BAGGING (5s_all_articles_BAGGING)	0.307	0.659	1.000	
RANDOM FOREST (5s_all_articles_rf)	0.297	0.184	0.086	1.000

Note 5.34: The table illustrates the correlation between the four supervised learning indicators based on all book reviews: five categories.

The classifiers which have been trained on the equalized corpus show a much more consistent picture. Only the *GLMENT* index seems to behave out of line at the beginning and at the end of the period. However, the remaining indices all show good results for the recession period and the two negative quarters with negative GDP growth (Figure 5:17).

Figure 5:17 - Classifiers trained on an equal number of book reviews: five classes (London)



Note 5.35: The figure illustrates the development of the eight supervised learning indicators, which have been trained by an equalized training corpus with five categories. As a test dataset, the London specific sub-corpus has been used.

The correlation analysis (Table 5:19) confirms this picture with high correlations among the majority of these indicators.

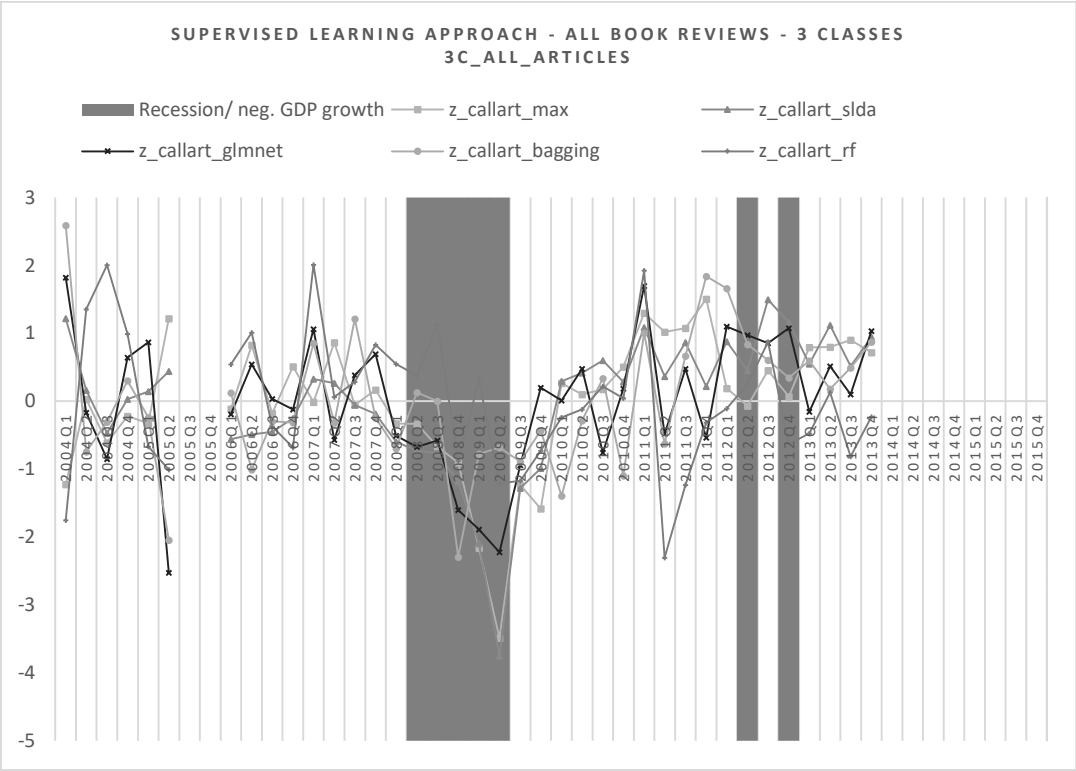
Table 5:19 - Correlation analysis - supervised learning approach (London) - 5 categories equal - equal number of reviews

	SVM (5s_eq_articles_*SVM*)	*MAXENT* (5s_eq_articles_max)	SLDA (5s_eq_articles_SLDA)	*GLMENT* (5s_eq_articles_*GLMENT*)	BOSSTING (5s_eq_articles_*BOOSTING*)	BAGGING (5s_eq_articles_BAGGING)	*RANDOM FOREST* (5s_eq_articles_rf)	Neural Net (5s_eq_articles_*NNET*)
SVM (5s_eq_articles_*SVM*)	1.000							
MAXENT (5s_eq_articles_max)	0.908	1.000						
SLDA (5s_eq_articles_SLDA)	0.943	0.919	1.000					
GLMENT (5s_eq_articles_*GLMENT*)	0.862	0.941	0.874	1.000				
BOSSTING (5s_eq_articles_*BOOSTING*)	0.798	0.718	0.809	0.692	1.000			
BAGGING (5s_eq_articles_BAGGING)	0.743	0.608	0.723	0.556	0.779	1.000		
RANDOM FOREST (5s_eq_articles_rf)	0.909	0.864	0.899	0.837	0.815	0.808	1.000	
Neural Net (5s_eq_articles_*NNET*)	0.835	0.916	0.878	0.915	0.750	0.563	0.843	1.000

Note 5.36: The table illustrates the correlation among the eight supervised learning indicators based on an equalized training corpus: five categories.

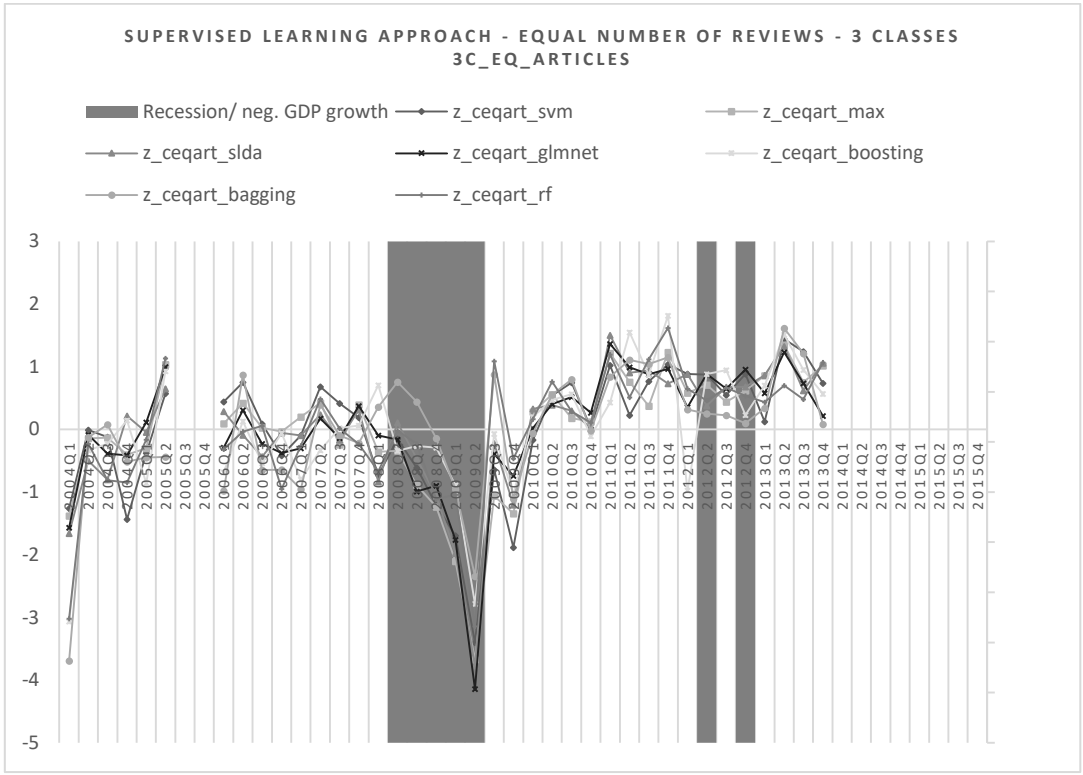
Figure 5:18 and Figure 5:19 display the results for the textual sentiment indices for London with three classes. It can be seen that those indicators which have been trained with all book reviews have improved upon their counterpart with five classes. However, compared to the equalized training set their result is still much more mixed.

Figure 5:18 - Classifiers trained on all book reviews: three classes (London)



Note 5.37: The figure illustrates the development of the five supervised learning indicators, which have been trained by the full training corpus with three categories. As a test dataset, the London specific sub-corpus has been used.

Figure 5:19 - Classifiers trained on an equal number of book reviews: three classes (London)



Note 5.38: The figure illustrates the development of the seven supervised learning indicators, which have been trained by an equalized training corpus with three categories. As a test dataset, the London specific sub-corpus has been used.

Table 5:20 - Correlation analysis supervised learning approach - (London) - 3 categories - all reviews

	MAXENT (3c_all_articles_max)	SLDA (3c_all_articles_SLDA)	*GLMENT* (3c_all_articles_*GLMENT*)	BAGGING (3c_all_articles_BAGGING)	*RANDOM FOREST* (3c_all_articles_rf)
MAXENT (3c_all_articles_max)	1.000				
SLDA (3c_all_articles_SLDA)	0.774	1.000			
GLMENT (3c_all_articles_*GLMENT*)	0.347	0.681	1.000		
BAGGING (3c_all_articles_BAGGING)	0.183	0.481	0.649	1.000	
RANDOM FOREST (3c_all_articles_rf)	0.101	0.063	0.211	0.083	1.000

Note 5.39: The table illustrates the correlation among the five supervised learning indicators based on all book reviews: three categories.

Table 5:21 - Correlation analysis -supervised learning approach - (London) - 3 categories - equal number of reviews

	SVM (3c_eq_articles_*SVM*)	*MAXENT* (3c_eq_articles_max)	SLDA (3c_eq_articles_SLDA)	*GLMENT* (3c_eq_articles_*GLMENT*)	*BOOSTING* (3c_eq_articles_*BOOSTING*)	BAGGING (3c_eq_articles_BAGGING)	*RANDOM FOREST* (3c_eq_articles_rf)
SVM (3c_eq_articles_*SVM*)	1.000						
MAXENT (3c_eq_articles_max)	0.927	1.000					
SLDA (3c_eq_articles_SLDA)	0.878	0.942	1.000				
GLMENT (3c_eq_articles_*GLMENT*)	0.897	0.947	0.937	1.000			
BOOSTING (3c_eq_articles_*BOOSTING*)	0.694	0.767	0.761	0.815	1.000		
BAGGING (3c_eq_articles_BAGGING)	0.684	0.644	0.698	0.722	0.843	1.000	
RANDOM FOREST (3c_eq_articles_rf)	0.804	0.806	0.855	0.853	0.799	0.806	1.000

Note 5.40: The table illustrates the correlation among the eight supervised learning indicators based on an equalized training corpus: three categories.

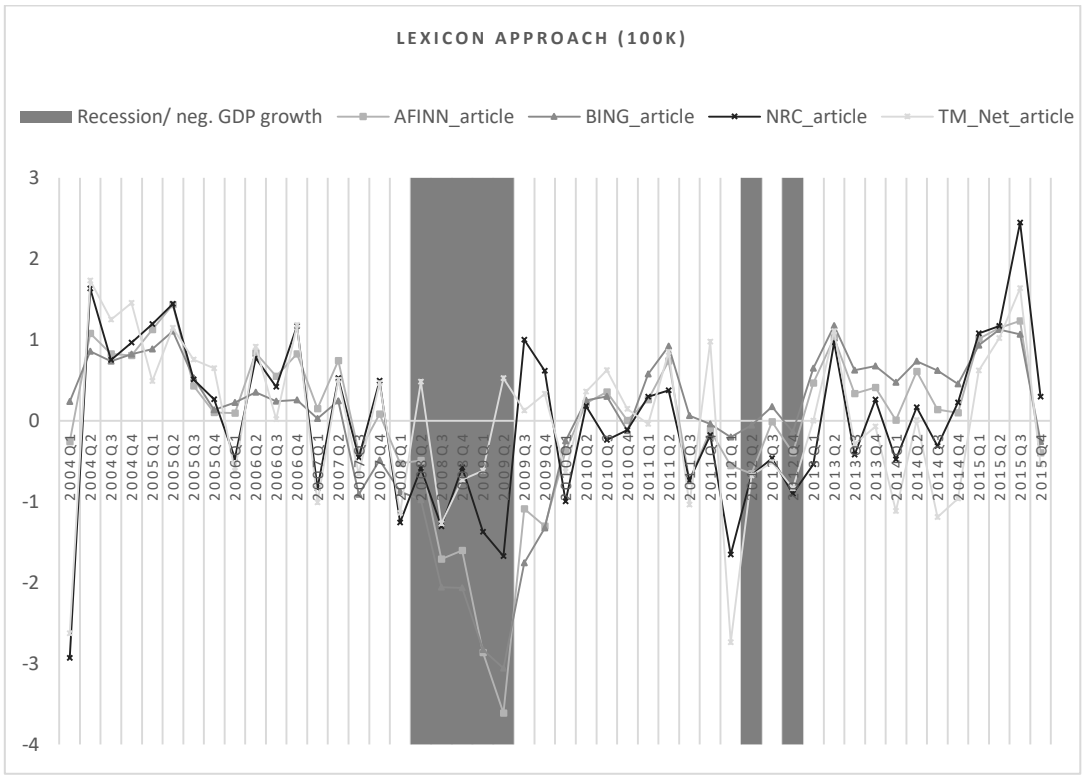
5.6.1.2.4 NEWSPAPERS WITH A CIRCULATION ABOVE 100,000 ISSUES

I created this sub-corpus to check whether newspapers with a broader coverage are more suitable to provide information about the commercial real estate market than its counterparts.

The results do not differ much from the previous sub-corpora. Therefore, I will illustrate the article charts as well as the corresponding correlation tables without any further comments.

LEXICON APPROACH

Figure 5:20 - Lexicon approach (100,000)



Note 5.41: The figure illustrates the development of the four different lexicon-based sentiment indicators on a quarterly base. The four algorithms mirror the sentiment for the 100,000 sub-corpus.

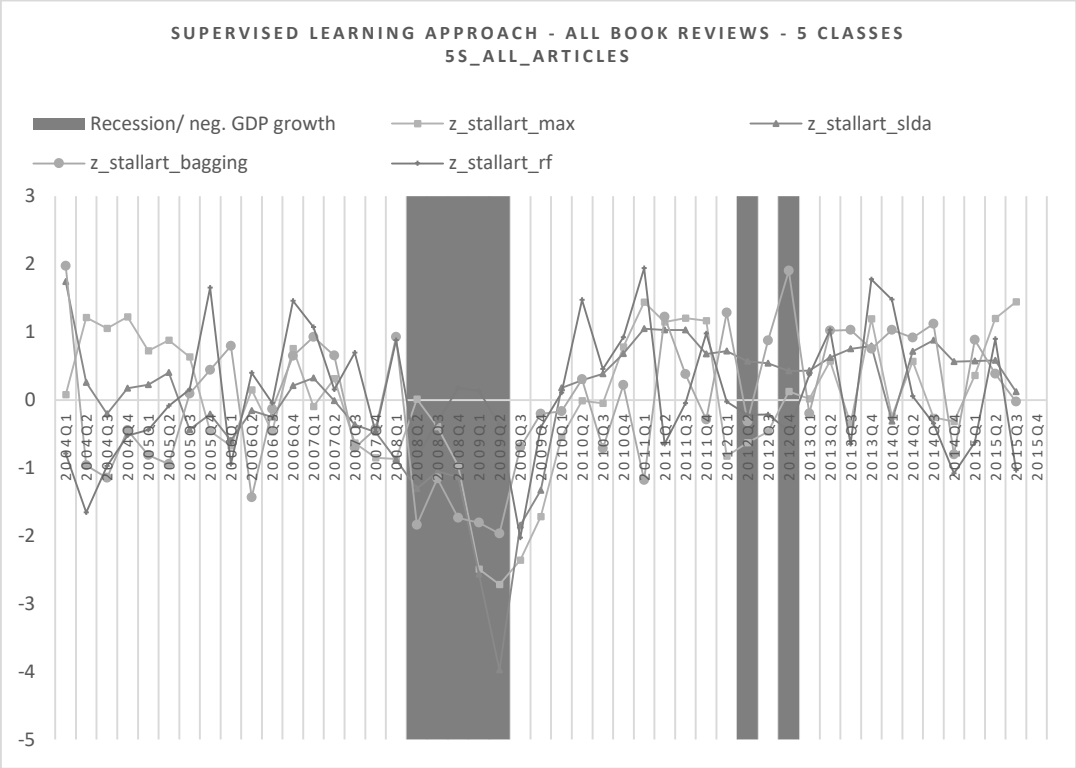
Table 5:22 - Correlation analysis - lexical indicators - (100,000)

	*AFINN*_article	*BING*_article	*NRC*_article	*TM*_Net_article
*AFINN*_article	1.000			
*BING*_article	0.939	1.000		
*NRC*_article	0.663	0.518	1.000	
*TM*_Net_article	0.456	0.318	0.812	1.000

Note 5.42: The table illustrates the correlation between the different sentiment indicators constructed by the lexicon approach.

SUPERVISED LEARNING APPROACH

Figure 5:21 - Classifiers trained on all book reviews: five classes (100,000)



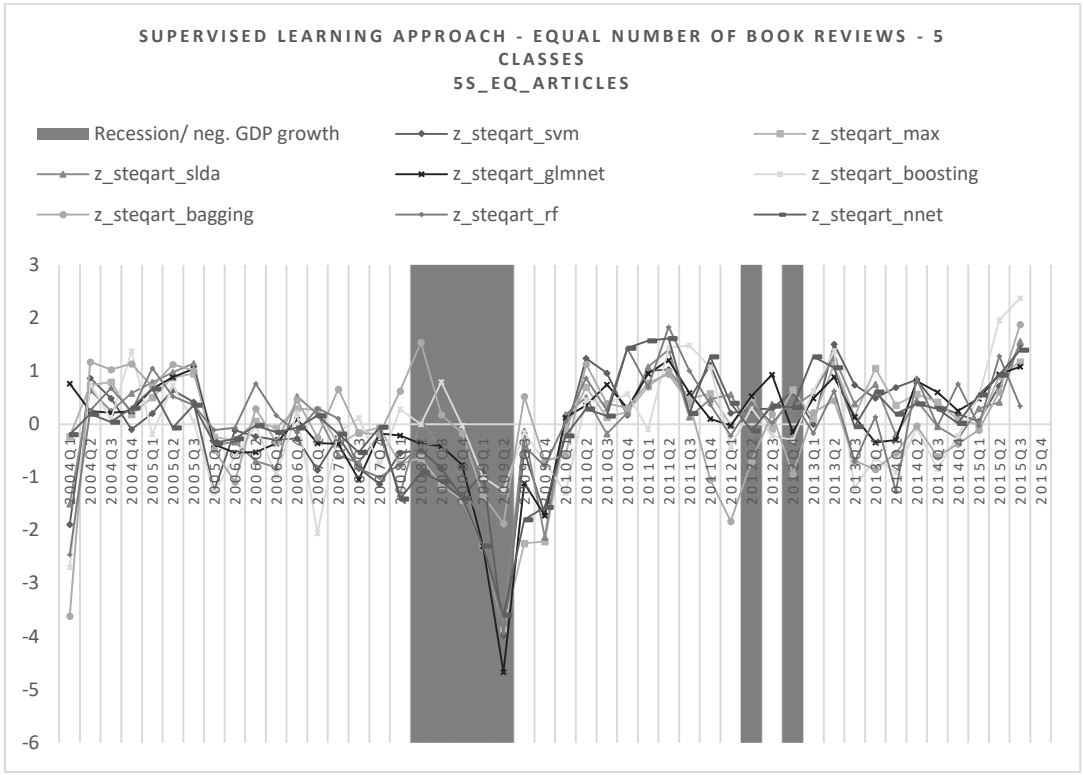
Note 5.43: The figure illustrates the development of the four supervised learning indicators, which have been trained by the full training corpus with five categories. As a test dataset, those news articles have been used, which were published by newspapers with more than 100,00 issues per day.

Table 5:23 - Correlation analysis - supervised learning approach - (100,000) - 5 categories - all reviews

	MAXENT (5s_all_articles_max)	SLDA (5s_all_articles_SLDA)	BAGGING (5s_all_articles_BAGGING)	*RANDOM FOREST* (5s_all_articles_rf)
MAXENT (5s_all_articles_max)	1.000			
SLDA (5s_all_articles_SLDA)	0.731	1.000		
BAGGING (5s_all_articles_BAGGING)	0.229	0.594	1.000	
RANDOM FOREST (5s_all_articles_rf)	0.176	0.158	0.123	1.000

Note 5.44: The table illustrates the correlation analysis between the four supervised learning algorithms trained on all book reviews with five categories.

Figure 5:22 - Classifiers trained on an equal number of book reviews: five classes (100,000)



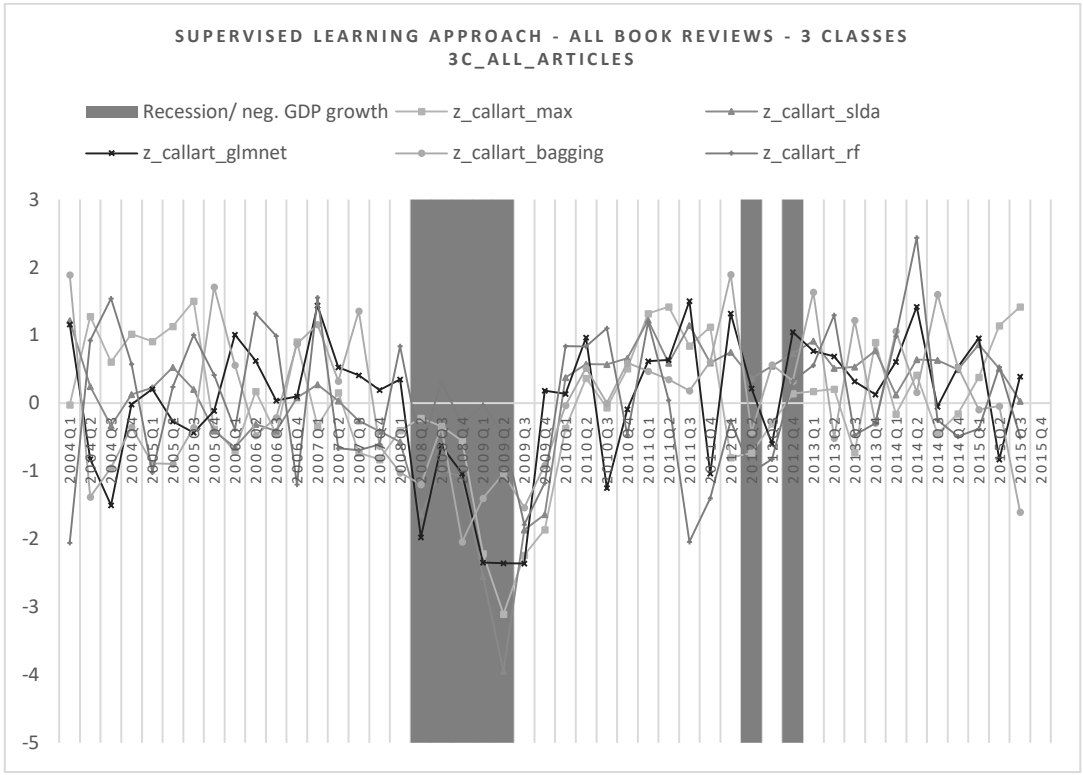
Note 5.45: The figure illustrates the development of the eight supervised learning indicators, which have been trained by an equalized training corpus with five categories. As a test dataset, those news articles have been used, which were published by newspapers with more than 100,00 issues per day.

Table 5:24 - Correlation analysis - supervised learning approach - (100,000) - 5 categories - equal number of reviews

	SVM (5s_eq_articles_*SVM*)	*MAXENT* (5s_eq_articles_max)	SLDA (5s_eq_articles_SLDA)	*GLMENT* (5s_eq_articles_*GLMENT*)	BOSSTING (5s_eq_articles_*BOOSTING*)	BAGGING (5s_eq_articles_BAGGING)	*RANDOM FOREST* (5s_eq_articles_rf)	Neural Net (5s_eq_articles_*NNET*)
SVM (5s_eq_articles_*SVM*)	1.000							
MAXENT (5s_eq_articles_max)	0.831	1.000						
SLDA (5s_eq_articles_SLDA)	0.844	0.860	1.000					
GLMENT (5s_eq_articles_*GLMENT*)	0.783	0.868	0.833	1.000				
BOSSTING (5s_eq_articles_*BOOSTING*)	0.574	0.402	0.529	0.420	1.000			
BAGGING (5s_eq_articles_BAGGING)	0.503	0.385	0.565	0.408	0.692	1.000		
RANDOM FOREST (5s_eq_articles_rf)	0.780	0.750	0.775	0.729	0.572	0.599	1.000	
Neural Net (5s_eq_articles_*NNET*)	0.788	0.893	0.844	0.843	0.433	0.321	0.752	1.000

Note 5.46: The table illustrates the correlation analysis among the eight supervised learning indicators based on an equal number of reviews with five categories.

Figure 5:23 - Classifiers trained on all book reviews: three classes (100,000)



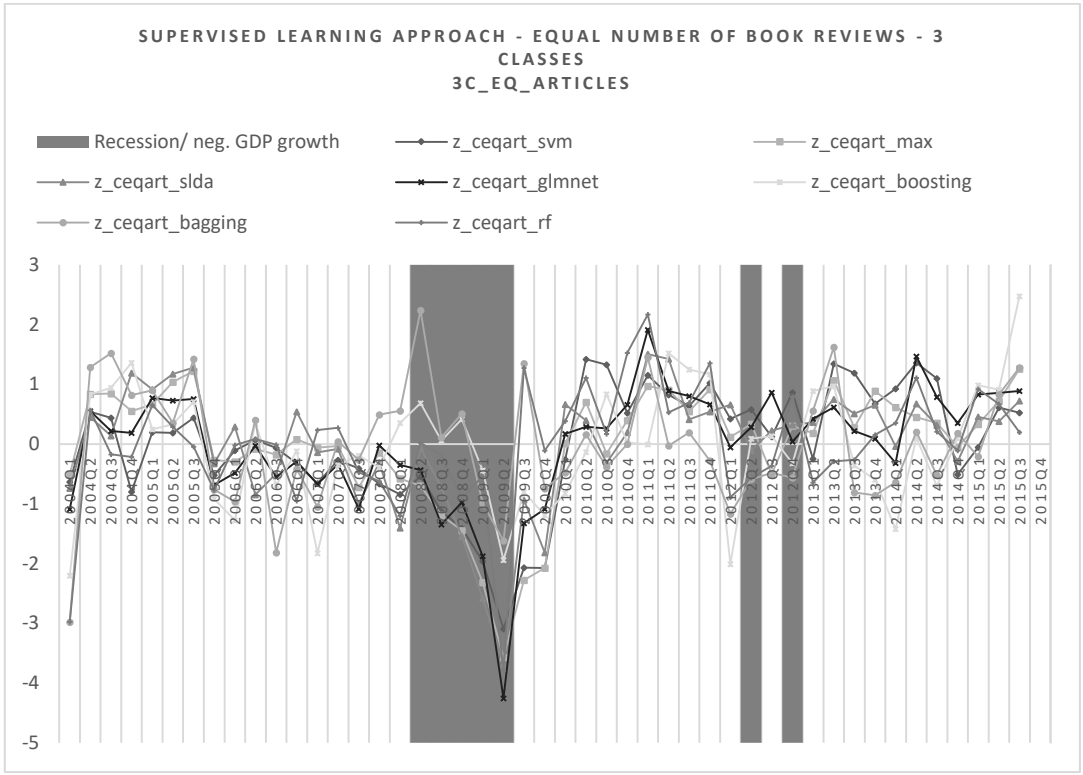
Note 5.47: The figure illustrates the development of the five supervised learning indicators, which have been trained by the full training corpus with three categories. As a test dataset, those news articles have been used, which were published by newspapers with more than 100,00 issues per day.

Table 5:25 - Correlation analysis -supervised learning approach - (100,000) - 3 categories - all reviews

	MAXENT (3c_all_articles_max)	SLDA (3c_all_articles_SLDA)	*GLMENT* (3c_all_articles_*GLMENT*)	BAGGING (3c_all_articles_BAGGING)	*RANDOM FOREST* (3c_all_articles_rf)
MAXENT (3c_all_articles_max)	1.000				
SLDA (3c_all_articles_SLDA)	0.728	1.000			
GLMENT (3c_all_articles_*GLMENT*)	0.346	0.676	1.000		
BAGGING (3c_all_articles_BAGGING)	0.088	0.535	0.564	1.000	
RANDOM FOREST (3c_all_articles_rf)	0.231	0.123	0.126	-0.023	1.000

Note 5.48: The table illustrates the correlation analysis among the five supervised learning indicators based on all book reviews with three categories.

Figure 5:24 - Classifiers trained on an equal number of book reviews: three classes (100,000)



Note 5.49: The figure illustrates the development of the seven supervised learning indicators, which have been trained by an equalized training corpus with three categories. As a test dataset, those news articles have been used, which were published by newspapers with more than 100,00 issues per day.

Table 5:26 - Correlation analysis - supervised learning approach - (100,000) - 3 categories - equal number of reviews

	SVM (3c_eq_articles_*SVM*)	*MAXENT* (3c_eq_articles_max)	SLDA (3c_eq_articles_SLDA)	*GLMENT* (3c_eq_articles_*GLMENT*)	*BOOSTING* (3c_eq_articles_*BOOSTING*)	BAGGING (3c_eq_articles_BAGGING)	*RANDOM FOREST* (3c_eq_articles_rf)
SVM (3c_eq_articles_*SVM*)	1.000						
MAXENT (3c_eq_articles_max)	0.865	1.000					
SLDA (3c_eq_articles_SLDA)	0.769	0.900	1.000				
GLMENT (3c_eq_articles_*GLMENT*)	0.815	0.869	0.867	1.000			
BOOSTING (3c_eq_articles_*BOOSTING*)	0.365	0.496	0.426	0.572	1.000		
BAGGING (3c_eq_articles_BAGGING)	0.182	0.329	0.360	0.435	0.688	1.000	
RANDOM FOREST (3c_eq_articles_rf)	0.549	0.524	0.592	0.690	0.399	0.458	1.000

Note 5.50: The table illustrates the correlation analysis among the seven supervised learning indicators based on an equal number of reviews with three categories.

5.6.1.2.5 FINANCIAL TIMES

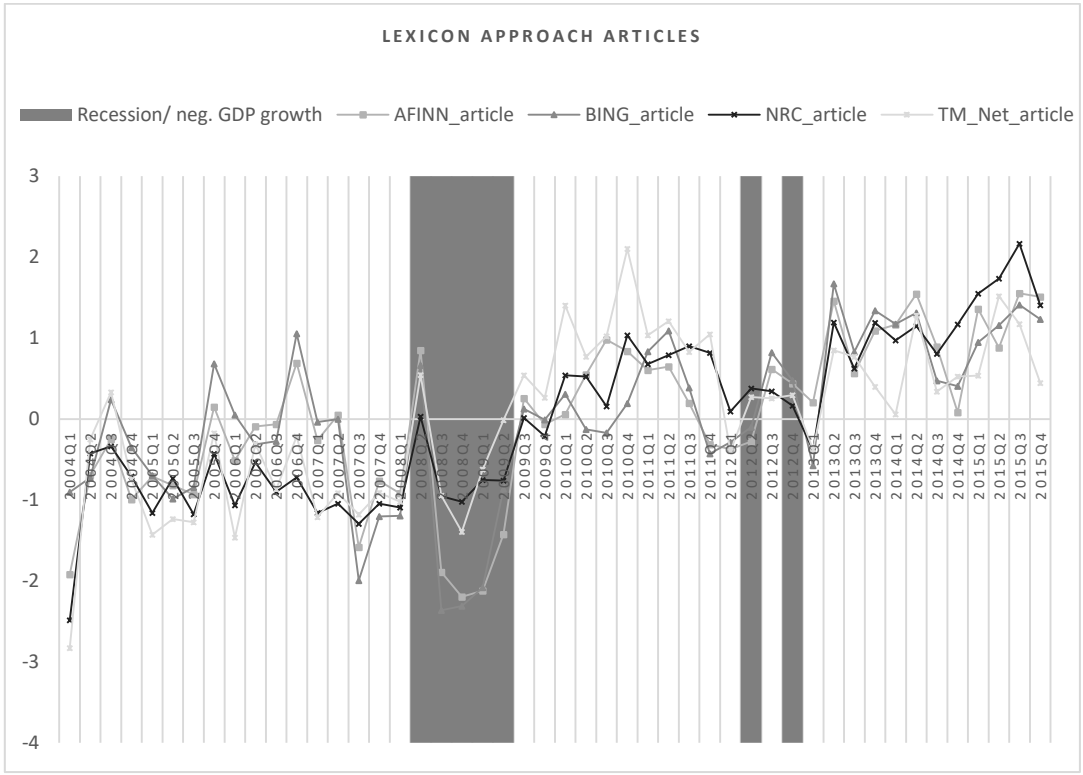
The *Financial Times* sub-corpus is based on 11,948 news articles. The reason why I have selected that specific newspaper is that I assume that real estate professionals read it on a daily basis. I further assume that the reader will be influenced by the content of the newspaper and therefore might change his or her behaviour. I am aware of the fact that this assumption would reduce the number of information sources down to one. Still, it is my belief that the newspaper has an excellent reputation and is widely read among professionals.

The graphical analysis reveals an entirely different picture than expected. It can be seen that the different classifiers are not in line as previously shown. One reason for this might be the fact that among all these different sub-corpora the total number of included articles is much lower. Another reason could be the fact that the *Financial Times* articles incorporate a much better description of the real estate market from a professional point of view, which incorporates multiple swings in the sentiment.

LEXICON APPROACH

Figure 5:25 illustrates the result of the lexical approach. During the primary recession period, it can be seen that the indicators reach their lowest values up to three to two quarters before the actual end of the recession. All indices do succeed, the expected development during the two quarters at the end of the observation period, by a minimum of one quarter.

Figure 5:25 - Lexicon approach (FT)



Note 5.51: The figure illustrates the development of the four different lexicon-based sentiment indicators on a quarterly base. The four algorithms mirror the sentiment for the Financial Times sub-corpus.

The correlation analysis shows a positive moderate to high correlation among the lexical sentiment indicators (Table 5:27).

Table 5:27 - Correlation analysis among the lexical indicators (FT)

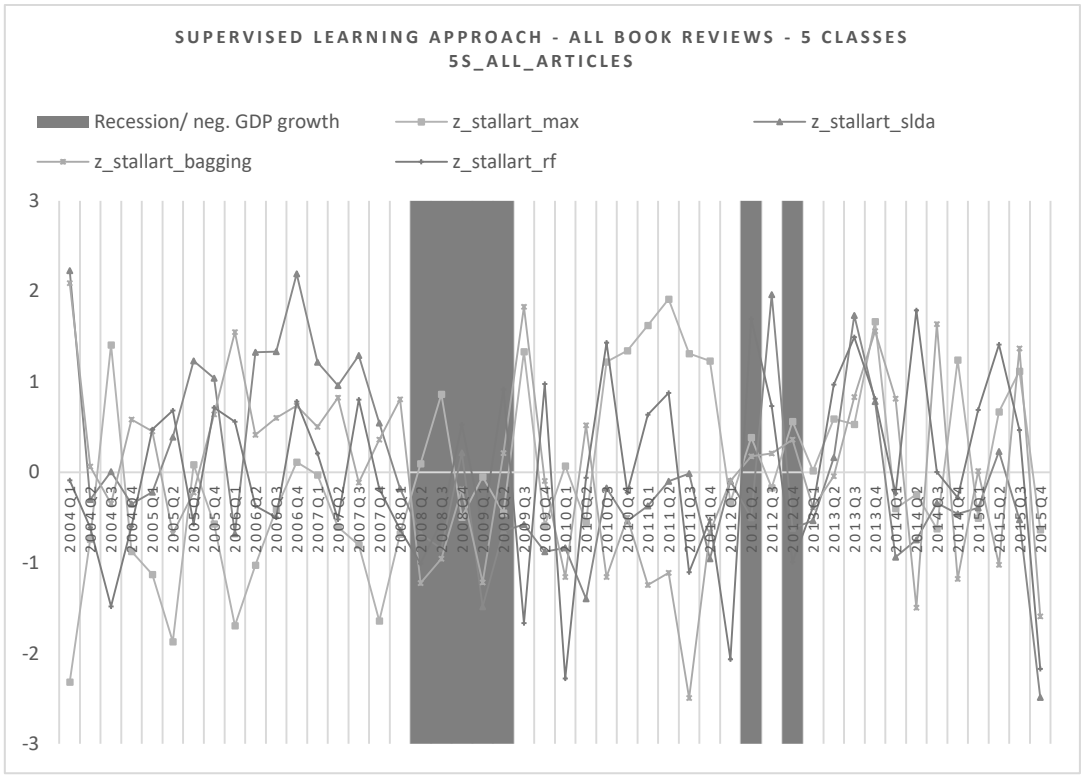
	*AFINN*_article	*BING*_article	*NRC*_article	*TM*_Net_article
*AFINN*_article	1.000			
*BING*_article	0.918	1.000		
*NRC*_article	0.805	0.738	1.000	
*TM*_Net_article	0.735	0.652	0.877	1.000

Note 5.52: The table illustrates the correlation between the different sentiment indicators constructed by the lexicon approach.

SUPERVISED LEARNING APPROACH

This picture becomes more chaotic over the analysis of the next four training sets with the different classes and different amounts of book reviews. Figure 5:26 reveals a similar picture as before. The use of the full set of reviews creates different qualities of classifiers. Over the course of the recession period, not all indicators show the negative development.

Figure 5:26 - Classifiers trained on all book reviews: five classes (FT)



Note 5.53: The figure illustrates the development of the four supervised learning indicators, which have been trained by the full training corpus with five categories. As a test dataset only, the Financial Times articles have been used.

The correlation analysis confirms this observation, with partly negative and low values (Table 5:28).

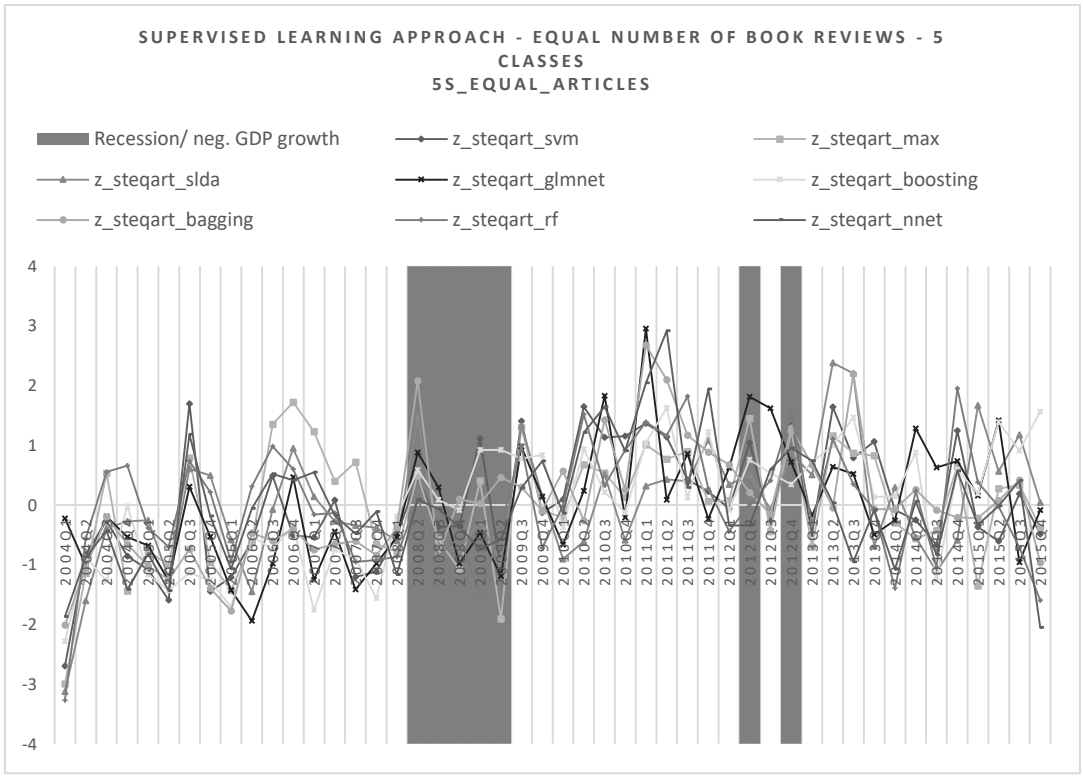
Table 5:28 - Correlation analysis - supervised learning approach - (FT) - 5 categories - all reviews

	MAXENT (5s_all_articles_max)	SLDA (5s_all_articles_SLDA)	BAGGING (5s_all_articles_BAGGING)	*RANDOM FOREST* (5s_all_articles_rf)
MAXENT (5s_all_articles_max)	1.000			
SLDA (5s_all_articles_SLDA)	-0.181	1.000		
BAGGING (5s_all_articles_BAGGING)	-0.355	0.388	1.000	
RANDOM FOREST (5s_all_articles_rf)	0.001	0.304	0.110	1.000

Note 5.54: The table illustrates the correlation analysis among the four supervised learning indicators based on all book reviews with five categories.

Similar to previous cases better results have been achieved with those classifiers which were trained on the equalized training corpus. Yet, even over the recession period, some indicators show a positive development (Figure 5:27).

Figure 5:27 - Classifiers trained on an equal number of book reviews: five classes (FT)



Note 5.55: The figure illustrates the development of the eight supervised learning indicators, which have been trained by an equalized training corpus with five categories. As a test dataset only, the Financial Times articles have been used.

Nevertheless, the results of the correlation analysis have slightly improved upon the full review training dataset (Table 5:29) with positive small to moderate correlations.

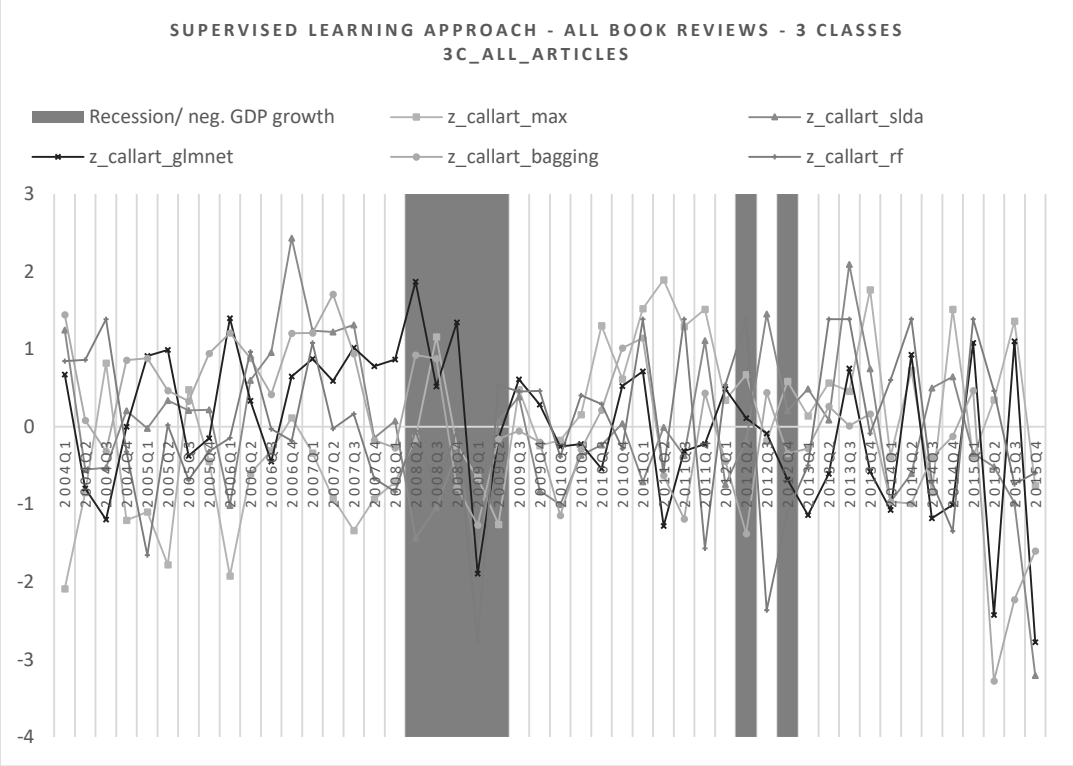
Table 5:29 - Correlation analysis -supervised learning approach - (FT) - 5 categories - equal number of reviews

	SVM (5s_eq_articles_*SVM*)	*MAXENT* (5s_eq_articles_max)	SLDA (5s_eq_articles_SLDA)	*GLMENT* (5s_eq_articles_*GLMENT*)	BOSSTING (5s_eq_articles_*BOOSTING*)	BAGGING (5s_eq_articles_BAGGING)	*RANDOM FOREST* (5s_eq_articles_rf)	Neural Net (5s_eq_articles_*NNET*)
SVM (5s_eq_articles_*SVM*)	1.000							
MAXENT (5s_eq_articles_max)	0.752	1.000						
SLDA (5s_eq_articles_SLDA)	0.518	0.580	1.000					
GLMENT (5s_eq_articles_*GLMENT*)	0.505	0.405	0.405	1.000				
BOSSTING (5s_eq_articles_*BOOSTING*)	0.568	0.336	0.528	0.496	1.000			
BAGGING (5s_eq_articles_BAGGING)	0.601	0.473	0.525	0.558	0.663	1.000		
RANDOM FOREST (5s_eq_articles_rf)	0.669	0.600	0.457	0.400	0.345	0.481	1.000	
Neural Net (5s_eq_articles_*NNET*)	0.616	0.539	0.361	0.392	0.365	0.574	0.647	1.000

Note 5.56: The table illustrates the correlation analysis among the eight supervised learning indicators based on an equal number of reviews with five categories.

Figure 5:28 illustrates the result of the full book review training corpus with three classes. It can be seen that some indicators (*GLMENT* or *MAXENT*) pick up the underlying market sentiment from the news articles. However, towards the end of the observation period, all indicators miss the two subsequent recession periods (except for the BAGGING indicator).

Figure 5:28 - Classifiers trained on all book reviews: three classes (FT)



Note 5.57: The figure illustrates the development of the five supervised learning indicators, which have been trained by the full training corpus with three categories. As a test dataset only, the Financial Times articles have been used.

The correlation coefficients are again better in comparison to the full training dataset using five different classes (Table 5:30). Yet, negative, as well as weak to moderate, positive correlations dominate this set of indicators.

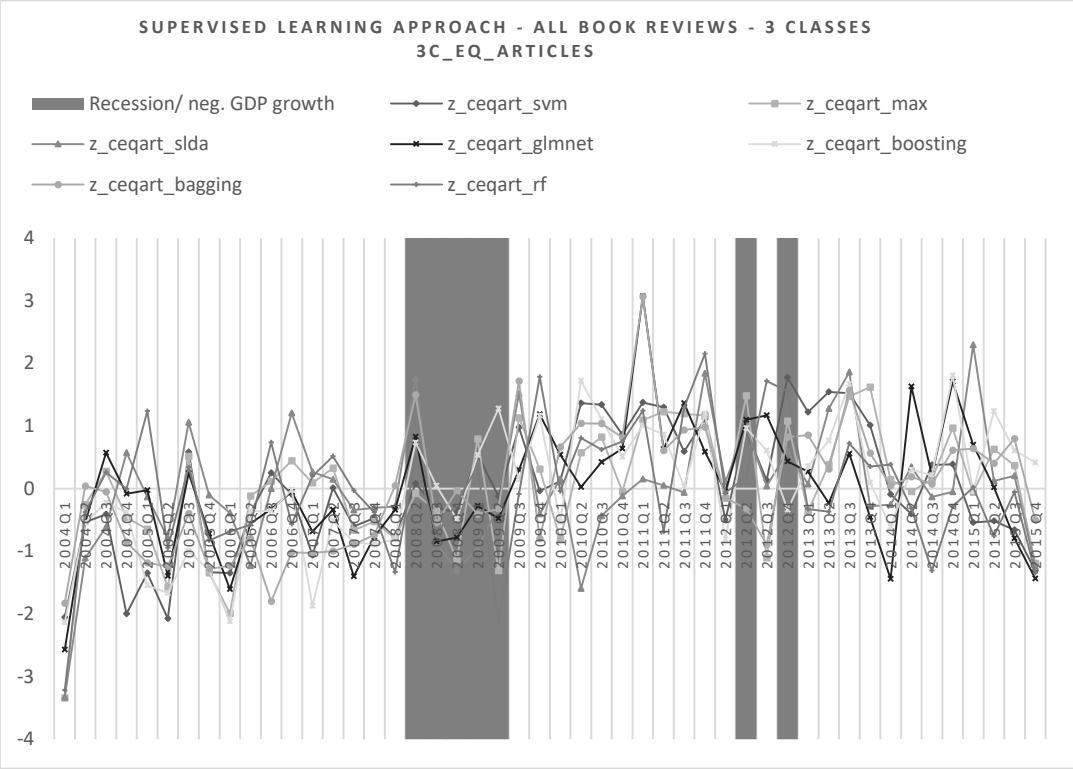
Table 5:30 - Correlation analysis - supervised learning approach - (FT) - 3 categories - all reviews

	MAXENT (3c_all_articles_max)	SLDA (3c_all_articles_SLDA)	GLMENT (3c_all_articles_GLMENT)	BAGGING (3c_all_articles_BAGGING)	*RANDOM FOREST* (3c_all_articles_rf)
MAXENT (3c_all_articles_max)	1.000				
SLDA (3c_all_articles_SLDA)	-0.025	1.000			
GLMENT (3c_all_articles_*GLMENT*)	-0.214	0.264	1.000		
BAGGING (3c_all_articles_BAGGING)	-0.289	0.488	0.526	1.000	
RANDOM FOREST (3c_all_articles_rf)	0.055	0.033	0.201	0.022	1.000

Note 5.58: The table illustrates the correlation analysis among the eight supervised learning indicators based on all book reviews: three categories.

The final set of indicators is produced by an equalized training data set using only three classes for the classification process. Figure 5:29 shows that the trend during the main recession period is more or less mirrored by the majority of indicators. In general, it can be observed that, compared to the previous sub-corpora, the indicators present a much more confusing picture. This could mean either that the number of articles in the construction process plays a more important role or that the extracted sentiment reacts to swings much more rapidly due to a small number of articles presenting the quarterly average value.

Figure 5:29 - Classifiers trained on an equal number of book reviews: three classes (FT)



Note 5.59: The figure illustrates the development of the seven supervised learning indicators, which have been trained by an equalized training corpus with three categories. As a test dataset only, the Financial Times articles have been used.

This improved behaviour of the indicators is further translated into higher correlation coefficients. In Table 5:31, the indicators are mainly moderately correlated.

Table 5:31 - Correlation analysis - supervised learning approach - (FT) - 3 categories - equal number of reviews

	SVM (3c_eq_articles_*SVM*)	*MAXENT* (3c_eq_articles_max)	SLDA (3c_eq_articles_SLDA)	*GLMENT* (3c_eq_articles_*GLMENT*)	*BOOSTING* (3c_eq_articles_*BOOSTING*)	BAGGING (3c_eq_articles_BAGGING)	*RANDOM FOREST* (3c_eq_articles_rf)
SVM (3c_eq_articles_*SVM*)	1.000						
MAXENT (3c_eq_articles_max)	0.759	1.000					
SLDA (3c_eq_articles_SLDA)	0.386	0.532	1.000				
GLMENT (3c_eq_articles_*GLMENT*)	0.576	0.615	0.493	1.000			
BOOSTING (3c_eq_articles_*BOOSTING*)	0.668	0.616	0.299	0.606	1.000		
BAGGING (3c_eq_articles_BAGGING)	0.683	0.605	0.406	0.624	0.623	1.000	
RANDOM FOREST (3c_eq_articles_rf)	0.495	0.586	0.425	0.572	0.303	0.343	1.000

Note 5.60: The table illustrates the correlation analysis among the eight supervised learning indicators based on an equal number of reviews with five categories.

5.6.1.2.6 SUMMARY

The graphical, as well as the correlation, analysis has revealed that a trained classifier on a biased training corpus (five or three categories with all reviews) produces fewer satisfying results. This has not only become clear in the no housing section but also in the other three sub-corpora (London, 100,000 and FT) as well. Classifiers which have used an equalized training dataset do not incorporate an initial bias toward the positive category.

The number of classes plays an essential role during the classification process. Fewer categories improve the graphical picture of the classifiers. The reason for this can be found in different methodologies of the classifiers. The separation of the indices and the corresponding sorting relies on less strict rules when only three classes are used.

It further seems that the number of articles in the test dataset matters as well. The smallest corpus of the *Financial Times* articles has produced diverse results for the different indicators. While I expected that the extracted sentiment would reveal a stronger insight into the actual market, it seems that the supervised learning indicators were unable to extract a sufficient amount of sentiment from the market, using the most recent financial crisis as an example. The topic was presented in nearly all newspapers at the same time, and therefore a better picture can be presented when a more significant share of the market – newspaper-wise – is used.

Based on the two performed analyses, a total of four textual sentiment indicators were selected for the implementation of the probit model. I have further considered the compiled $F - score$ in the initial analysis. The four selected indicators reached an $F - score$ of above 0.5.

Given the good performance in the graphical analysis, the Maximum Entropy indicator ($3c\_all\_MAXENT$) based on all training documents with three categories was selected.[32] Further, the Support Vector Machine indicator ($3c\_eq\_SVM$), the Maximum Entropy indicator ($3c\_eq\_MAXENT$)[33] and the *RANDOM FOREST* indicator ($3c\_eq\_RANDOM\ FOREST$) based on the equalized training dataset will be used for the analysis. This set should provide a full picture of the sentiment

In addition, the four textual sentiment indicators based on the lexical approaches, *BING*, *AFINN*, *NRC* and *TM*, will be tested in a probit model. These indicators proved in the previous chapter that they are able to extract the sentiment with the help of the underlying word lists.

[32] From now on also referred to as *MAXENT* II.
[33] From now on also referred to as *MAXENT* I.

The probit model analysis is used to compare further these simple indicators with the somewhat complicated supervised learning indicators.

Table 5:32 summarizes the correlation among the selected indicators. The correlations are mainly moderate to a strong, which specifies that the selected indices will present a similar picture of the extracted sentiment.

Table 5:32 - Correlation between leading indicators

	AFINN	BING	NRC	TM	SVM	Maximum Entropy (1)	RANDOM FOREST	Maximum Entropy (2)
AFINN	1.000							
BING	0.934	1.000						
NRC	0.695	0.627	1.000					
TM	0.596	0.517	0.882	1.000				
Support Vector Machine	0.778	0.728	0.445	0.298	1.000			
Maximum Entropy (1)	0.817	0.814	0.610	0.481	0.738	1.000		
RANDOM FOREST	0.674	0.568	0.706	0.548	0.614	0.637	1.000	
Maximum Entropy (2)	0.827	0.787	0.588	0.462	0.757	0.804	0.615	1.000

Note 5.61: The table illustrates the correlation between the eight selected leading indicators. In general, the correlation among these indicators is moderate to high, indicating that the indicators share a common trend.

5.6.1.3 CORRELATION ANALYSIS BETWEEN THE RICS U.K. COMMERCIAL MARKET SURVEY AND THE TEXTUAL SENTIMENT INDICATORS

In this section, I try to justify the use and the quality of the constructed sentiment indicators. Here I like to address the issue that the applied methodology in the above-presented analysis is unknown in quality, especially when it comes to the supervised learning algorithms. The reason for this is that the classifiers are based on a training dataset, which is unknown in structure, content and sentiment. Therefore, the quality of the sentiment indicators remains hidden. This obviously does not apply for the lexicon-based classifiers. To justify further the use of the method, a correlation analysis between the textual sentiment indicators and the RICS U.K. commercial market survey is performed. Ideally, the series will show a positive correlation, indicating a common ground of information. As has been described in the literature review, sentiment extracted from interviews or surveys has been proven to be superior compared to indirect sentiment proxies. However, I have also described the disadvantages of the use of a survey-based measure, which become especially prominent in the absence of such an indicator.

For the U.K., the RICS publishes a regular property survey-based sentiment indicator on a quarterly level. The survey is structured into various categories. Two outputs are the general Sales and Rental Levels and the Office Sales and Rent Levels in London for the next quarter. Both series reach back until 1998. Survey participants express their expectations about the market development for the upcoming quarter. The opinion of all participants is then aggregated and summarized in a single value.

Since the series is only available on a quarterly level, we need to convert the RICS values into a monthly series. The indicators have been standardized in order to be comparable to the textual sentiment indicators.

On the side of the textual sentiment indicators, I will apply the eight selected sentiment indicators (AFINN, BING, NRC, TM, SVM, MAXENT (equal articles), MAXENT (all articles) and Random Forrest). Starting with the *AFINN* model, Table 5:33 illustrates the correlation between the five different *AFINN* models and the two RICS survey measures. All values range between 0.389 and 0.641, which indicates a moderate to a strong positive relationship.

Looking at the values in more detail, it can be seen that the all articles indicator scores higher for the more general London survey measure. This is also true for the other sub-corpora, except for the 100,000. The highest correlation is achieved by the London specific index.

For the *BING* model, the scores range between 0.425 and 0.722 (Table 5:33), which mirrors a moderate to strong positive correlation. Similar to before, the higher correlations are achieved by the RICS general sales and rental level expectations for the London market. The London indicator has again the highest correlation to the survey measures.

Both the *NRC* and the *TM* indicators behave in a similar fashion. For the *NRC* indicators the correlation range between 0.189 and 0.524. For the *TM* measures, the coefficients range between 0.046 and 0.463. As expected, the results are weaker in comparison to the other two lexicon measures.

The *SVM* method has produced correlation coefficients between 0.063 and 0.512. The correlation remains weak to moderate. With essentially no to a moderate correlation, again the more general survey measure reveals higher correlations. Similar to before the 100,000 measure has produced the best result.

The correlation coefficients for the *MAXENT* I models are lower in comparison. The values range from 0.087 to 0.578 (Table 5:33). Therefore, the correlation between the *MAXENT* I model and the RICS measures can be described as weak to moderate. Different to before, this time the highest correlations are achieved by the all articles indicators.

This pattern remains for the second MAXENT measure. The correlation coefficients range between 0.015 and 0.442. As expected, the results are slightly weaker in direct comparison to the former *MAXENT* indicator.

The weakest overall result is produced by the *Random Forrest* measure. There is essentially no correlation. Only the all articles indicator produces a weak relationship with bot RICS series.

Table 5:33 - Correlation table between the AFINN, BING and MAXENT I indicators and the U.K. RICS survey measures

	Sales & rental levels in London	Office sales & rent levels in London
AFINN (all)	0.574	0.565
AFINN (no housing)	0.634	0.612
AFINN (London)	0.641	0.621
AFINN (100,000)	0.589	0.604
AFINN (FT)	0.416	0.389
BING (all)	0.652	0.627
BING (no housing)	0.721	0.680
BING (London)	0.722	0.683
BING (100,000)	0.711	0.691
BING (FT)	0.462	0.425
NRC (all)	0.362	0.361
NRC (no housing)	0.524	0.503
NRC (London)	0.397	0.391
NRC (100,000)	0.189	0.216
NRC (FT)	0.251	0.198
TM (all)	0.260	0.264
TM (no housing)	0.463	0.429
TM (London)	0.334	0.325
TM (100,000)	0.046	0.077
TM (FT)	0.118	0.097
SVM equal articles (all)	0.497	0.461
SVM equal articles (no housing)	0.344	0.307
SVM equal articles (London)	0.443	0.431
SVM equal articles (100,000)	0.512	0.485
SVM equal articles (FT)	0.065	0.063
MAXENT equal articles (all)	0.578	0.559
MAXENT equal articles (no housing)	0.416	0.370
MAXENT equal articles (London)	0.489	0.477
MAXENT equal articles (100,000)	0.541	0.521
MAXENT equal articles (FT)	0.100	0.087
MAXENT all articles (all)	0.442	0.421
MAXENT all articles (no housing)	0.389	0.352
MAXENT all articles (London)	0.430	0.429
MAXENT all articles (100,000)	0.315	0.329
MAXENT all articles (FT)	0.031	0.015
Random Forrest equal articles (all)	0.332	0.321
Random Forrest equal articles (no housing)	0.168	0.153
Random Forrest equal articles (London)	0.228	0.246
Random Forrest equal articles (100,000)	0.179	0.208
Random Forrest equal articles (FT)	0.005	0.039

Note 5.62: The table above reports the correlation between the five different AFINN, BING and MAXENT I sentiment measures and the two U.K. RICS direct sentiment measures.

Overall the correlation analysis reveals that the textual sentiment indicators have a weak to moderate positive correlation to one of the leading sentiment indicators of the U.K. In some cases, as for the BING method, the correlation is strong. This underlines the qualities of these newly constructed indicators. Different from the survey-based measures, the textual sentiment indicators mirror the market in its current stage. At least the lexicon approach models are relatively easy to construct and provide a good indication of the market movement.

5.6.1.4 PROBIT MODEL

In this section, I use the extracted textual sentiment indicators within a probit framework. As described in the first study of this thesis, different approaches have been developed over the years. While sentiment proxies share the characteristics of the macroeconomic indicators, textual sentiment indicators are different in their nature. The main difference is the ability of the textual indicators to reflect on the current situation more or less isochronically.

The following analysis is quite extensive and will bring all the previous parts together. As described above I will not use all developed textual sentiment indicators, but eight in total. This central section will use two *MSCI* series, which I have converted into a binary growth rate. The series is the *MSCI* all properties and the *MSCI* all offices leading indicators.

Each of the two dependent variables will be tested against the eight sentiment indicators. The section is separated into the analysis of the five sub-corpora (*all articles, no housing, London, 100,000 and the FT sub-corpus*). In the beginning, I will present the descriptive statistics of the used variables and the results of the Augmented Dickey-Fuller Test. The third part will show the regression results regarding the two dependent variables.

Next, for the standard regression outcomes, I have provided the pseudo-R-squared value to evaluate the quality of the different indicators. Furthermore, I have checked the classification score with similar sensitivity and specificity values, which indicate how well the textual sentiment indicators have performed in the two classes of the binary variable. Finally, I have used the *Hosmer-Lemeshow* χ^2 test and the Receiver Operating Characteristic (*ROC*) curve to evaluate the quality of the residuals.

Each section ends with a simple in-sample forecast as well as a forecast test for the occurring turning points of the dependent variables.

5.6.1.4.1 SUB-CORPUS I: ALL ARTICLES

This first sub-corpus uses all collected articles. In comparison to the other four corpora, this one can be seen as the least specific since it includes those articles which contain housing or residential related terms. In addition, the number of included newspapers is higher than in the subsequent tries.

Table 5:34 - Summary of statistics (all articles)

Variable	Obs	Mean	Std. Dev.	Min	Max
All assets all properties (*MSCI*_change of growth rate)	158	0.297	0.459	0.000	1.000
All assets all offices (*MSCI*_change of growth rate)	158	0.272	0.446	0.000	1.000
AFINN	144	0.000	1.000	-3.579	1.803
BING	144	0.000	1.000	-2.914	1.941
NRC	144	0.000	1.000	-8.470	2.001
TM	144	0.000	1.000	-7.881	2.358
SVM (equal articles)	144	0.000	1.000	-4.070	1.934
MAXENT (equal articles)	144	0.000	1.000	-4.512	1.766
RANDOM FOREST (equal articles)	144	0.000	1.000	-7.174	1.667
MAXENT (all articles)	144	0.000	1.000	-3.685	2.160

Note 5.63: The table illustrates the summary of statistics.

Table 5:34 illustrates the descriptive statistics for the different variables. The two dependent variables have been converted into a binary series with 0 and 1; 1 for those instances where negative growth was observed. All series are given in monthly observations. The two sets of textual sentiment indicators (lexicon and machine learning approaches) have been standardized with a mean of 0 and a standard deviation of 1. Different to the dependent variables, only 144 observations between January 2004 and December 2015 are recorded for the textual indicators.

None of the ten variables shows any sign of a unit root. The test statistics of the Augmented Dickey-Fuller (ADF) tests have all been higher than the critical value at the 1% level. The difference in the number of the observations in Table 5:35 results from the fact that I had used lagged variables during the ADF test. The number of lags was determined by the Akaike Information Criteria (AIC), as stated in Table 5:36. The uses variables are, therefore, assumed to be stationary.

Table 5:35 - Augmented Dickey-Fuller Test (all articles)

Variable	Test statistics	1% critical value	5% critical value	10% critical value	Obs.
All assets all properties (*MSCI*_change of growth rate)	-3.568	-3.491	-2.886	-2.576	157
All assets all offices (*MSCI*_change of growth rate)	-4.046	-3.491	-2.886	-2.576	157
AFINN	-4.583	-3.496	-2.887	-2.577	142
BING	-3.424	-3.496	-2.887	-2.577	142
NRC	-4.656	-3.496	-2.887	-2.577	141
TM	-3.846	-3.497	-2.887	-2.577	139
SVM (equal articles)	-5.935	-3.496	-2.887	-2.577	142
MAXENT (equal articles)	-3.954	-3.496	-2.887	-2.577	142
RANDOM FOREST (equal articles)	-7.813	-3.496	-2.887	-2.577	142
MAXENT (all articles)	-4.876	-3.496	-2.887	-2.577	142

Note 5.64: The table illustrates the results of the Augmented Dickey-Fuller Test. All test-statistics are above the critical values at a 1% level.

5.6.1.4.1.1 PROBIT MODEL RESULTS (ALL ARTICLES)

Since the ADF test has not revealed any unit root, I run the eight different probit models against the first dependent variable: the converted *MSCI* all properties growth rate. Table 5:36 illustrates the individual regression results.

First, it can be seen that all coefficients are highly significant at the 1% level and that they have a negative impact on the dependent variable. This result confirms my expectations since the conversion of the dependent variable leads to a mirrored image of the actual market movement. While the market experienced a negative development over the course of the financial crisis, in the probit framework, those negative events are now marked as positive. However, since those negative events do only represent a minor share in comparison to the overall series, the textual sentiment indicators are required to influence the dependent variable negatively.

As a measure of goodness of fit McFadden's pseudo-R-squared is presented. Since the R-squared value cannot be interpreted similarly to the R-squared value of an OLS regression, they should be treated with caution. Values around 0.2 can be seen as reasonable. Only three of the eight models show values within that range. The *AFINN* indicator (0.195) and the *MAXENT* series (0.179) are only outperformed by the *BING* series (0.281). All the remaining models show lower values.

Table 5:36 - Probit results: MSCI - all assets - all properties (all articles)

Dependent variable *MSCI* all assets all properties		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	Description	*AFINN*_Articles	*BING*_Articles	*NRC*_Articles	*TM*_Articles	Support Vector Machine	Maximum Entropy (1)	RANDOM FOREST	Maximum Entropy (2)
l.z_*AFINN*_article	Standardized values for the lexicon approach with the *AFINN* lexicon	-0.706***							
		[0.135]							
l.z_*BING*_article	Standardized values for the lexicon approach with the *BING* lexicon		-0.898***						
			[0.149]						
l2.z_*NRC*_article	Standardized values for the lexicon approach with the *NRC* lexicon			-0.301***					
				[0.100]					
l4.z_tm_article	Standardized values for the lexicon approach with the *TM* lexicon				-0.309***				
					[0.103]				
l.z_ceqart_*SVM*	Standardized values for the *SVM* algorithm based on the equalized training corpus with 3 categories					-0.515***			
						[0.128]			
l.z_ceqart_max	Standardized values for the *MAXENT* algorithm based on the equalized training corpus with 3 categories						-0.679***		
							[0.134]		
l.z_ceqart_rf	Standardized values for the RF algorithm based on the equalized training corpus with 3 categories							-0.327***	
								[0.105]	
l.z_callart_max	Standardized values for the *MAXENT* algorithm based on the full training corpus with 3 categories								-0.538***
									[0.127]
Constant		-0.624***	-0.673***	-0.576***	-0.557***	-0.594***	-0.624***	-0.576***	-0.603***
		[0.122]	[0.129]	[0.113]	[0.113]	[0.117]	[0.121]	[0.114]	[0.118]
Observations		144	144	144	144	144	144	144	144
Log-likelihood		-69.93	-62.47	-82.57	-83.43	-77.51	-71.39	-82.09	-76.56
LR Chi2		33.99	48.91	8.703	8.724	18.82	31.06	9.671	20.72
Number of lags		1	1	2	4	1	1	1	1
pseudo-R-squared		0.195	0.281	0.050	0.050	0.108	0.179	0.056	0.119
AIC		143.862	128.941	169.144	170.865	159.027	146.788	168.177	157.128
BIC		149.802	134.881	175.084	176.805	164.967	152.728	174.116	163.067
Correctly classified (%)		79.17	81.25	70.83	69.44	73.61	78.47	71.53	75.00
Sensitivity		42.86	54.76	2.38	0.00	23.81	42.86	4.76	30.95
Specificity		94.12	92.16	99.02	99.01	94.12	93.14	99.02	93.14
Hosmer-Lemeshow χ^2		4.340	8.600	11.710	7.250	9.930	7.260	6.720	3.660
Prob > χ^2		0.825	0.377	0.165	0.506	0.270	0.509	0.568	0.887
area under Receiver Operating Characteristic (*ROC*) curve		0.787	0.830	0.752	0.725	0.703	0.772	0.711	0.723

Standard errors in brackets; *** p<0.01, ** p<0.05, * p<0.1

Note 5.65: The table illustrates the probit results for the MSCI, all assets, all properties series. It can be seen that all textual sentiment indicators, who have extracted the sentiment from the full news-corpus, remain highly significant at a 1% level. The lexicon approaches (AFINN and BING) do outperform the supervised learning measures according to the pseudo-R-squared value.

To elaborate on these results, I ran three additional diagnostic tests. The first concerns the classification of the values. The overall rate of correct classification for the *BING* model is estimated to be 81.25, with 54.76% of the average weight group correctly classified (specificity) and 92.16% of the low weight group correctly classified (sensitivity). Classification is sensitive to the relative sizes of each component group and always favours classification into the larger group. This phenomenon is evident here since only a minor number of observations of the dependent variable falls into the normal weight group. As a cut-off point for the classification, I have used 0.5.

The *AFINN* and the Maximum Entropy Model I show similar results, with an equally good distribution of the observations into either one of the categories. Models 3, 4 and 7, on the other hand, fail to sort the observations accordingly and over-sort one of the categories.

Next, I performed the *Hosmer-Lemeshow* χ^2 test. The test can also be seen as a measure of goodness of fit. Values with high positive figures and a corresponding p-value of above 0.05 indicate that the models predicted probabilities that broadly match the event rates. The corresponding p-values for the eight models are all above 0.5, which indicates that all models provide acceptable results.

The last diagnostic test is the analysis of the area of the Receiver Operating Characteristic (*ROC*) curve or the C-statistic. Values of around 0.7 are seen as acceptable. Values of around 0.5 indicate that the observations are sorted into either one of the categories more or less randomly. The results in Table 5:36 show that all models produce satisfying results, with values above this threshold. Again, the *BING* measure produces the best result with an area under the *ROC* curve of 0.83.

To summarize, three models seem to be capable of explaining the dependent variable. Furthermore, the observation I made earlier in this chapter, that the machine learning indicators do not perform as good as the lexicon indicators, prevails. It seems that the extraction of the sentiment with word lists is not only more straightforward but also more efficient in comparison to the text classification with the *Amazon* book reviews.

Since in this thesis I try to focus on the commercial real estate market, I have tried to select only those news articles which tend to discuss commercial real estate. To test if the sentiment is more directed towards this side of the market, the second dependent variable is more specific and only uses the modified *MSCI* all offices growth rate.

Table 5:37 illustrates the results of the eight probit models. I am able to report that the results are very similar to the previous analysis. Again, all models produce highly significant coefficients at the 1% level, with both the constant and the textual sentiment indicator being negative. Different from the previous results, the lag structure of the individual indicators has changed slightly. While in Table 5:36 only two indicators (*NRC* and *TM*) had more than one lag, now six of the eight have at least two lags. That indicates that the leading series precedes market development. Given that, the dependent variable is now a bit more directed towards the specific market. The reader should not forget that the underlying basis for this analysis uses all articles, which naturally incorporates some noise.

Starting the discussion of the results again with McFadden's pseudo-R-squared it can be observed that the *BING* model again outperforms the remaining models. Compared to the previous result, the value has further increased and is now at 0.345. The *AFINN* and the Maximum Entropy I models come second and third with corresponding pseudo-R-squared values of 0.243 and 0.221. The remaining models fail to generate values within an acceptable range of 0.2.

Compared to the results in Table 5:36 most of the classification values have improved. For the *BING* model, the overall value of correctly classified observations is now 83.33. The sensitivity score has slightly decreased (54.05%), but specificity (93.46%) has gone up in comparison.

It is only worth mentioning that the area under the *ROC* curve has also been improved by the *BING* model and that all remaining models still produce values close to and above 0.7.

To summarize: the idea that commercial real estate related articles carry more market-relevant information can be seen as proven despite the variety in quality differences among the different models. Setting a stronger focus on the commercial real estate side has led to more significant results when the dependent variable is more related to the CRE market.

Table 5:37 - Probit results: MSCI - all assets - all offices (all articles)

Dependent Variable: *MSCI* all assets - office properties		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	Description	*AFINN*_Articles	*BING*_Articles	*NRC*_Articles	*TM*_Articles	Support Vector Machine	Maximum Entropy (1)	RANDOM FOREST	Maximum Entropy (2)
l2.z_*AFINN*_article	Standardized values for the lexicon approach with the *AFINN* lexicon	-0.794***							
		[0.142]							
l2.z_*BING*_article	Standardized values for the lexicon approach with the *BING* lexicon		-1.025***						
			[0.164]						
l2.z_*NRC*_article	Standardized values for the lexicon approach with the *NRC* lexicon			-0.326***					
				[0.101]					
l3.z_tm_article	Standardized values for the lexicon approach with the *TM* lexicon				-0.322***				
					[0.105]				
l.z_ceqart_*SVM*	Standardized values for the *SVM* algorithm based on the equalized training corpus with 3 categories					-0.560***			
						[0.133]			
l2.z_ceqart_max	Standardized values for the *MAXENT* algorithm based on the equalized training corpus with 3 categories						-0.756***		
							[0.139]		
l.z_ceqart_rf	Standardized values for the RF algorithm based on the equalized training corpus with 3 categories							-0.345***	
								[0.106]	
l2.z_callart_max	Standardized values for the *MAXENT* algorithm based on the full training corpus with 3 categories								-0.630***
									[0.137]
Constant		-0.784***	-0.873***	-0.691***	-0.667***	-0.720***	-0.787***	-0.690***	-0.752***
		[0.131]	[0.144]	[0.117]	[0.116]	[0.122]	[0.130]	[0.117]	[0.125]
Observations		144	144	144	144	144	144	144	144
Log-likelihood		-62.13	-53.73	-77.08	-78.49	-71.38	-63.93	-76.77	-69.27
LR Chi2		39.84	56.65	9.954	9.22	21.36	36.26	10.56	25.58
Number of lags		2	2	2	3	1	2	1	2
pseudo-R-squared		0.243	0.345	0.061	0.056	0.130	0.221	0.064	0.156
AIC		128.268	111.464	158.159	160.981	146.755	131.856	157.550	142.534
BIC		134.208	117.404	164.098	166.920	152.695	137.795	163.490	148.474
Correctly classified (%)		82.64	83.33	74.31	72.92	77.08	81.94	75.00	78.47
Sensitivity		43.24	54.05	2.70	0.00	27.03	45.95	5.41	32.43
Specificity		96.26	93.46	99.07	99.06	94.39	94.39	99.07	94.39
Hosmer-Lemeshow χ^2		2.980	8.420	13.540	8.400	13.090	5.910	6.200	5.700
Prob > χ^2		0.936	0.394	0.094	0.395	0.109	0.657	0.625	0.681
area under Receiver Operating Characteristic (*ROC*) curve		0.830	0.870	0.774	0.733	0.726	0.823	0.729	0.766

Standard errors in brackets; *** p<0.01, ** p<0.05, * p<0.1

Note 5.66: The table illustrates the probit results for the MSCI, all assets, all offices series. It can be seen that all textual sentiment indicators, who have extracted the sentiment from the full news-corpus, remain highly significant at a 1% level. The lexicon approach BING does outperform the remaining indicators according to the pseudo-R-squared value.

5.6.1.4.1.2 PREDICTIONS (ALL ARTICLES)

In this part, I provide the predicted probability graphs for the two sets of the textual sentiment indicators for both dependent variables.

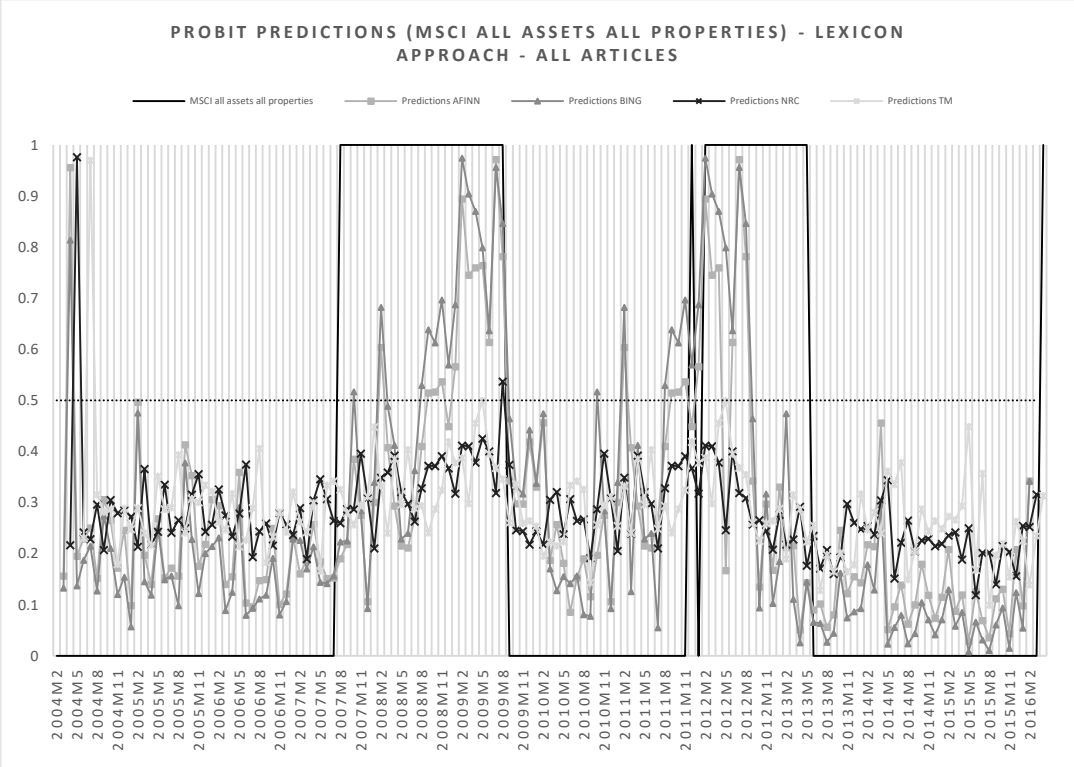
Both Figure 5:30 and Figure 5:31 show the probabilities for the *MSCI* all properties series. It can be seen that over the course of the 144 months, three periods show a negative growth rate: (i) between September 2007 and August 2009; (ii) December 2011; and (iii) between March 2012 and May 2015.

Over the course of the first seven months, all four sentiment indicators peak at between 0.8 and just shy of 1. However, the leading series remain in the below 0.5 area, and therefore below the baseline, afterwards. When the first period with negative growth sets in (2007M8) the *AFINN* and the *BING* indicator climb over the baseline towards the negative area. Both series remain in the negative area over the course of the recession period. While this development has been successively, the turn towards the more positive growth area is more or less instantaneous.

During the second longest period of negative growth, the *BING* indicator was adopted by August 2011, which is eight months before the actual negative growth was recorded. The reason for this could be that authors were still quite sensitive to a possible negative development in the market, and might have fallen back into the language of the financial crisis. The *BING* series does not mirror the full negative period until the end of May 2015, which indicates a change in the language of the authors.

During the period after May 2015, all indicators act accordingly and remain in the expected area.

Figure 5:30 - Prediction of the MSCI all properties series - lexicon approach (all articles)



Note 5.67: The figure illustrates the probit predictions of the four lexicon measures, which have extracted the sentiment from the full corpus, for the MSCI all assets all properties series.

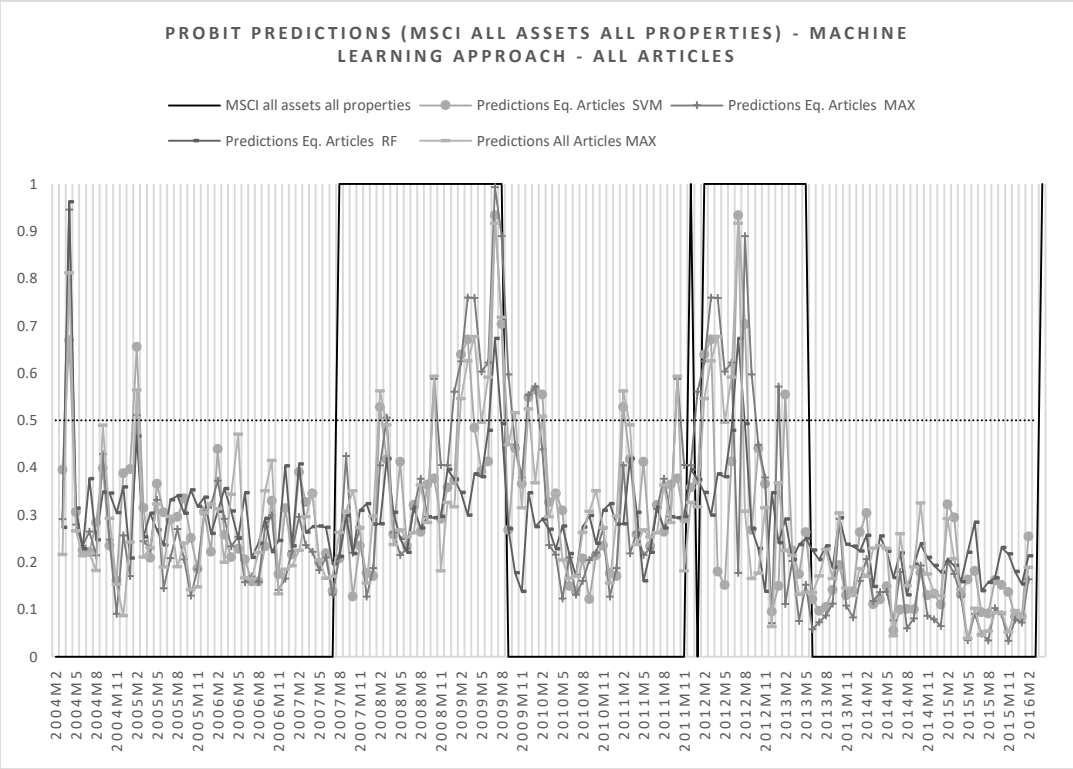
Figure 5:31 illustrates the predicted probabilities for the four machine learning algorithms. As the results section has shown, their overall performance is less satisfying. Different from the graphs of the lexicon approaches, all four indicators seem to be much closer together, only the *RANDOM FOREST* series shows some contradictory results in various stages.

Similar to the previous figure all four indicators show a peak in the first 14 months. They also seem to fail to pick up the trend and show some extreme changes when the first negative growth period sets in. Towards the end of the financial crisis, all indicators drop back into the expected area with lower probabilities.

As the second-long negative growth period occurs, some indicators rise more or less instantaneously above the baseline. However, the observed variation is much more extreme with the indicator switching between the two states. Similar to the lexicon approach, the four machine learning series drop back into the below baseline area way before the end of the event.

During the period after May 2015, again all indicators act accordingly and remain in the expected area.

Figure 5:31 - Prediction of the MSCI all properties series - machine learning approach (all articles)



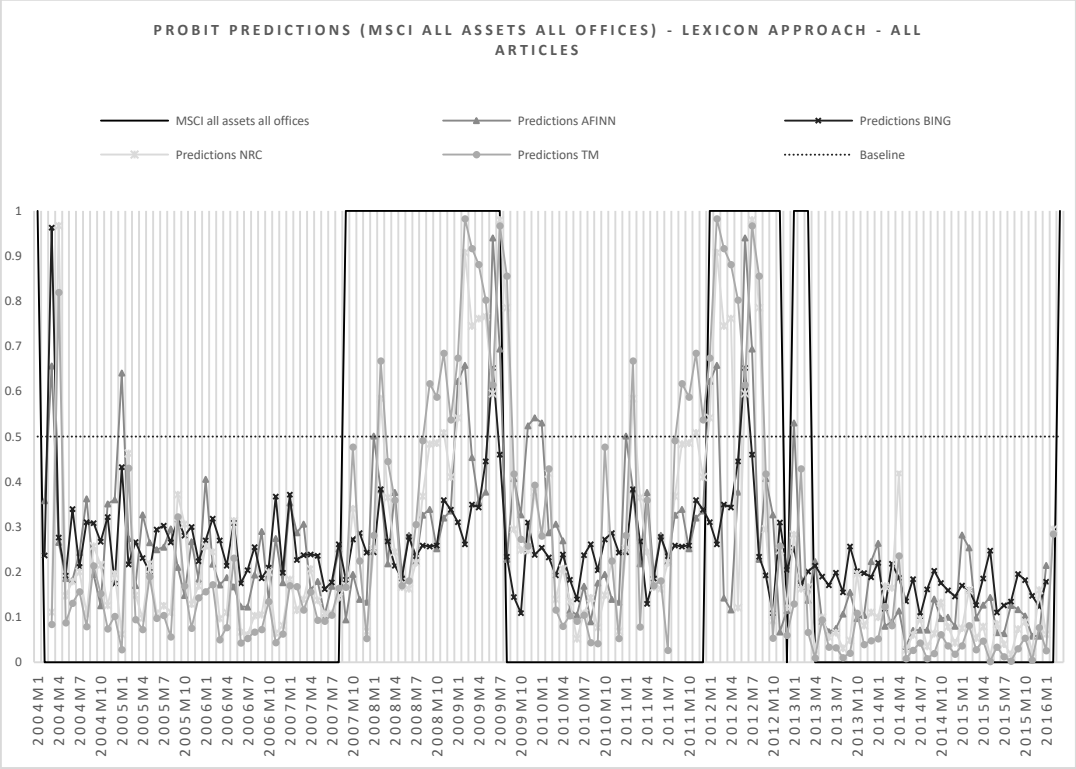
Note 5.68: The figure illustrates the probit predictions of the four supervised learning measures, which have extracted the sentiment from the full corpus, for the MSCI all assets all properties series.

Figure 5:32 and Figure 5:33 illustrate the probabilities for the *MSCI* all office series. Different from the all properties series (Figure 5:30 and Figure 5:31), the *MSCI* all office series shows a one-month gap in the second period of negative growth. December 2012 reveals no negative growth.

Figure 5:32 illustrates the results for the lexicon approach indicators. While the *BING* series achieved the best results in the regression part, its probability scores do not resemble the overall trend of the dependent variable. During the financial crisis, the series only peaks once towards the end. In the second phase of negative development, the indicator also oversteps the baseline once in the middle. This does not resemble the quality of the good results.

Entirely different from the previous results is the behaviour of the *TM* and *NRC* series. They are now much more able to follow the overall trend of the dependent variable. While the *TM* indicator is able to pick up the negative development during the financial crisis and in the second larger period of observed negativity, it also shows some variation inbetween those periods.

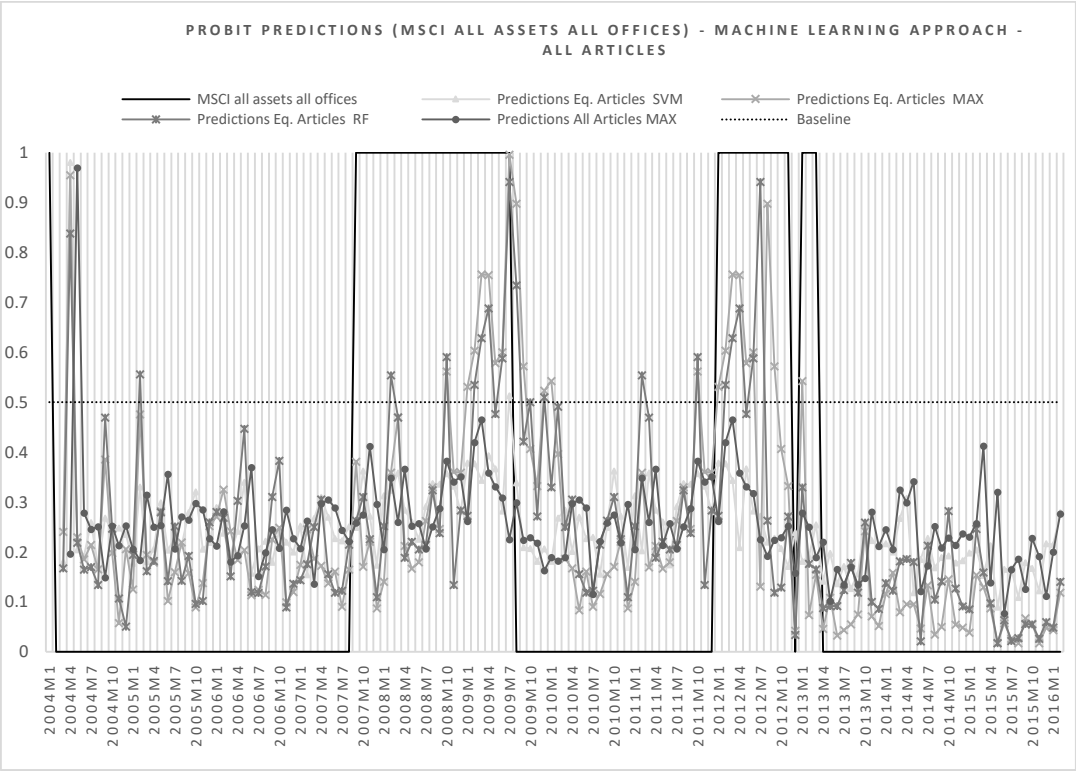
Figure 5:32 - Prediction of the MSCI all offices series - lexicon approach (all articles)



Note 5.69: The figure illustrates the probit predictions of the four lexicon measures, which have extracted the sentiment from the full corpus, for the MSCI all assets all offices series.

For the machine learning indicators (Figure 5:33), the picture is somewhat similar in the fact that the *RANDOM FOREST* index shows a strong resemblance to the dependent variable. This is surprising compared to the relatively low probit result quality. However, the *BING* series is, as expected, also picking up the negative phases; yet it is much stronger in the second period from December 2011 onwards. What is positive is that the last period after the second longer negative growth period is characterized by a stable and below the baseline behaviour for all textual sentiment indicators.

Figure 5:33 - Predictions of the MSCI all offices series - machine learning approach (all articles)



Note 5.70: The figure illustrates the probit predictions of the four supervised learning measures, which have extracted the sentiment from the full corpus, for the MSCI all assets all offices series.

To summarize, different to the previous two (Figure 5:30 and Figure 5:31), the three well-performing indicators (*AFINN*, *BING*, Maximum Entropy I) fail to mirror the dependent variable to the same extent. One reason could be that those obviously significant indicators extracted a much more directed sentiment from the articles. Unfortunately, this sentiment is unable to produce adequate probability results. Maybe a more directed underlying dependent variable could improve upon the results.

The second conclusion which can be drawn from this first result is the fact that all figures show, especially to the end of the financial crisis, a peak in their development. During the graphical analysis, I made a similar observation. As I stated earlier, I believe that the authors of those articles use the first signs of improvement within the market to summarize past developments and advise market participants to handle things with caution. Afterwards, the language changes entirely and the positive description of expected developments push the sentiment up.

5.6.1.4.1.3 DIEBOLD MARIANO TEST (ALL ARTICLES)

The previous tests have revealed that the *BING* model outperforms the other models. In this section, I will estimate the forecast significance of the different models in relation to the assumed superior model. The Diebold Mariano Test can be seen as an in-sample test.

For this purpose, I performed the Diebold Mariano Test as proposed in Diebold and Mariano (1995). The test determines a measure of predictive accuracy given an actual series. It uses two competing predictions against one another. I decided to report the mean squared error (MSE) as the measure of forecast accuracy. The DM test calculates a number of measures for predictive accuracy, to test the null hypothesis of equal accuracy.

S (1) is the measure which calculates the mean difference between the loss criteria for the two predictions. In this case, it is zero when there is no difference between the two predictions. Due to the structure of the test, the long-run estimate of the variance of the difference is used. Therefore, the test can also be described as quite data hungry and I have not restricted any testing periods, but used the full sample of the predicted values.

Table 5:38 illustrates the results for the eight different models from the overall corpus section. As stated earlier, the models are all tested against the *BING* model. Therefore, each line refers to the *BING* model.

The results suggest that *BING* with its lexicon approach produces the best prediction of the dependent variable in comparison. The mean MSE (0.137) is at least 0.018 times smaller than the next model (*AFINN*). Surprising is that the *MAXENT* (equal articles) model also computes a reasonably small MSE, yet the S (1) statistic is insignificant at the 10% level.

Table 5:38 - Diebold-Mariano Test - MSCI all properties all assets (all articles)

	MSE	Difference	S (1)	p-value
BING	0.137			
AFINN	0.155	-0.018	-2.105	0.035
NRC	0.191	-0.053	-2.426	0.015
TM	0.193	-0.058	-2.430	0.015
SVM (equal articles)	0.179	-0.043	-2.236	0.025
MAXENT (equal articles)	0.158	-0.022	-1.486	0.137
RANDOM FOREST (equal articles)	0.190	-0.053	-2.279	0.023
MAXENT (all articles)	0.176	-0.039	-2.421	0.016

Note 5.71: The table illustrates the results of the Diebold-Mariano Test for the MSCI all properties all assets series, for those indicators, which have extracted the sentiment from the full news-corpus. The BING series has been used as a reference for the test and all remaining series are evaluated against it.

Table 5:39 illustrates the DM test results for the *MSCI* all office series. The picture regarding the superiority of the *BING* indicator remains unchanged. Again, *BING* outperforms the other seven indicators and shows the lowest MSE.

Table 5:39 - Diebold Mariano Test - MSCI all properties all offices (all articles)

	MSE	Difference	S (1)	p-value
BING	0.118			
AFINN	0.136	-0.018	-1.966	0.049
NRC	0.175	-0.057	-3.121	0.002
TM	0.178	-0.060	-2.600	0.009
SVM (equal articles)	0.162	-0.044	-3.465	0.001
MAXENT (equal articles)	0.140	-0.022	-1.377	0.169
RANDOM FOREST (equal articles)	0.175	-0.057	-3.178	0.002
MAXENT (all articles)	0.158	-0.041	-2.467	0.014

Note 5.72: The table illustrates the results of the Diebold-Mariano Test for the MSCI all properties all offices series, for those indicators, which have extracted the sentiment from the full news-corpus. The BING series has been used as a reference for the test and all remaining series are evaluated against it.

5.6.1.4.1.4 TURNING POINTS (ALL ARTICLES)

In this section, I perform an in-sample forecast to predict the turning points of the dependent variables. For the *MSCI* all properties series these are 2009m8, 2012m1 and 2013m5. The first actual turning point in 2007m7 cannot be tested due to the lack of data variation. The third turning point in 2011m1 is only one period and is, therefore, ignored. I run an out-of-sample forecast, where I have developed the individual models until three months before the occurrence of the turning point and then predicted the next six periods.

The models are compared against each other and against the naïve approach, where the last observation is assumed to be the value of the next period. In addition, I have only used the *AFINN*, the *BING* and the *MAXENT* (equal articles) indicators in this exercise, since they have produced the most significant and promising results in the above-presented analyses.

Table 5:40 - Forecast evaluation for the three turning points of the MSCI all properties series (all articles)

	First turning point 2009m8			Second turning point 2012m1			Third turning point 2013m5		
Measures of forecast accuracy	AFINN	BING	*MAXENT* (equal articles)	AFINN	BING	*MAXENT* (equal articles)	AFINN	BING	*MAXENT* (equal articles)
Mean forecast error	-0.104	-0.202	-0.175	0.046	-0.130	0.166	0.376	0.435	0.371
Mean absolute error	0.258	0.264	0.332	0.379	0.337	0.433	0.471	0.479	0.463
Mean squared error	0.085	0.119	0.155	0.172	0.200	0.203	0.366	0.421	0.354
Root mean squared error	0.292	0.345	0.394	0.415	0.447	0.450	0.605	0.649	0.595
Theil's U1	0.214	0.237	0.279	0.308	0.322	0.367	0.717	0.828	0.704
Theil's U2	0.413	0.488	0.557	0.509	0.548	0.551	0.856	0.917	0.842
C-statistic	-0.829	-0.761	-0.689	-0.740	-0.699	-0.695	-0.266	-0.157	-0.290

Note 5.73: The table evaluates the forecast results for the three turning points of the MSCI all properties series. In this analysis, only the three best performing textual sentiment measures, based on the full corpus, have been applied. For each of the turning points, the forecast has been performed individually. All series have been estimated until three months before the occurrence of the turning point and then the next six periods have been predicted.

Table 5:40 illustrates the measure of forecast accuracy for the three selected models, based on the overall news corpus predicting the *MSCI* all properties series. Starting with the mean forecast error, it can be seen that higher values are achieved by the models for the third turning point, while the second period of interest produces the lowest values in comparison. Comparing the three models with each other, the *AFINN* approach has the smallest difference to zero, where over and underestimations of the actual values would cancel each other out. All models have a negative mean forecast error for the first turning point period, indicating an overreaction of the forecast values. For the other two periods, those signs swap, except for the *BING* induced model during the second period of interest.

For the mean squared error, small values are desired. The measure can be used to compare different methods with each other. Unfortunately, it can be seen that only the MSE of the *AFINN* model for the first turning point has a relatively small value of 0.085.

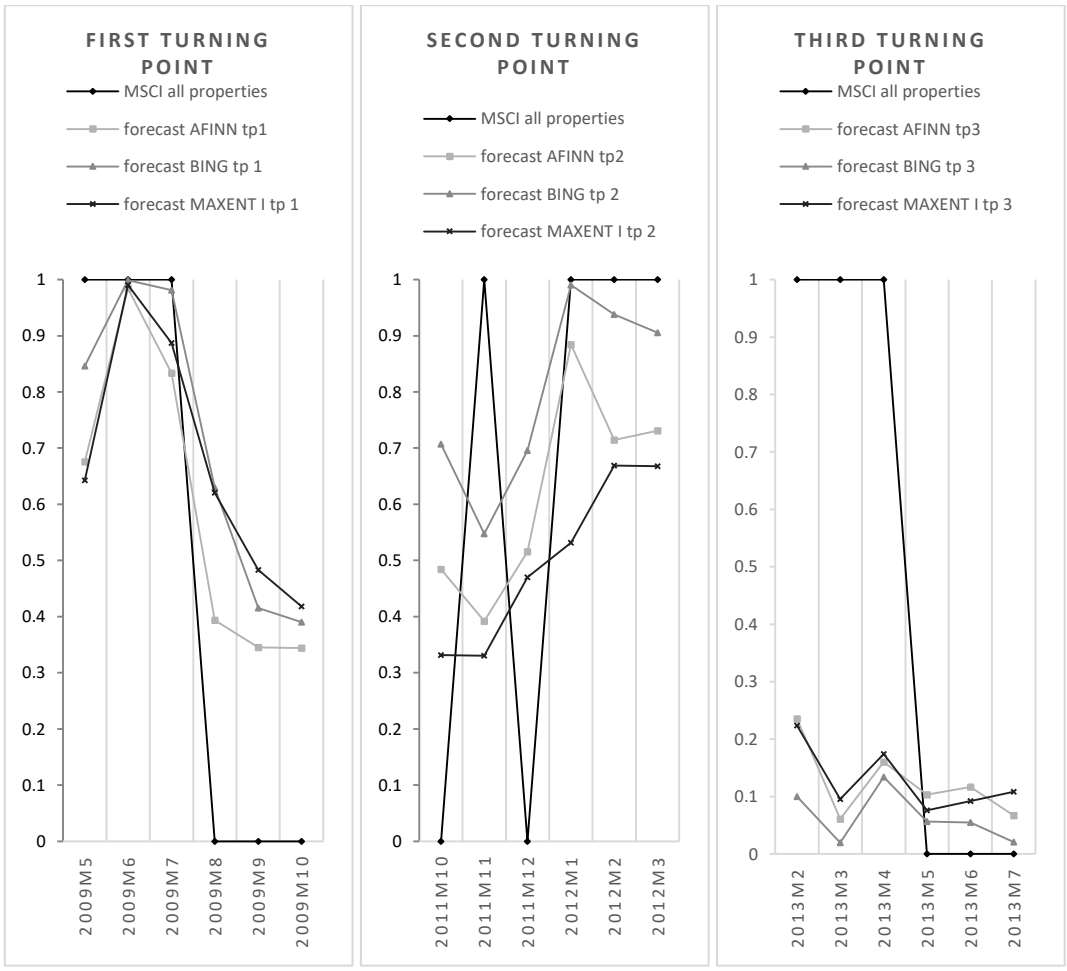
Theil's U1 evaluates the prediction performance. Values closer to zero than 1 are preferred. For the first period, all models show results below 0.3. Unfortunately, all models show a sharp

increase in the second and third period, which means that their prediction performance is rather bad.

The last two forecast measures Theil's U2 and the C-statistic show that all models outperform the naïve approach. The values of the Theil's U2 measure are smaller than one, and the negative values of the C-statistic confirm to this original picture.

Figure 5:34 illustrates the predicted turning points by the three different models.

Figure 5:34 - Turning point predictions MSCI all properties (all articles)



Note 5.74: The three graphs above illustrate the development of the forecast of the textual sentiment indicators during turning points. The dependent variable in this analysis is the MSCI all properties all assets series.

To conclude, sentiment induced models are able to improve upon the naïve approach. Comparing the above-presented results with the results from the DM test, it is surprising that the *BING* approach is not the best model in this analysis.

Table 5:41 shows the forecast evaluation for the *MSCI* all office series for the three different methods. Similar to the previous result, the mean forecast error has negative values over the first turning point period and swaps the signs in the subsequent periods. Only the *BING* approach (–0.040) shows for the second turning point a negative sign, with the smallest value for all methods and periods, meaning that nearly all errors cancel each other out.

The results for the mean squared error have been improved in direct comparison to Table 5:40, with values below 0.2 for the first and second turning points. Over the second period, the *BING* model is able to reveal the smallest value in comparison. Yet, for the first and third period, the *AFINN* sentiment induced model produces much smaller values.

Regarding Theil's U1 it can be seen that the values increase from period to period, with the exemption of the *BING* model (0.237), which shows its smallest value during the second turning point.

The last two remaining forecast measures again show that all models outperform the naïve approach. The values of Theil's U2 measure are smaller than one, and the negative values of the C-statistic confirm this original picture.

Table 5:41 - Forecast evaluation for the three turning points MSCI all offices (all articles)

	First turning point 2009m8			Second turning point 2012m1			Third turning point 2013m4		
Measures of forecast accuracy	AFINN	BING	*MAXENT* (equal articles)	AFINN	BING	*MAXENT* (equal articles)	AFINN	BING	*MAXENT* (equal articles)
Mean forecast error	-0.160	-0.294	-0.188	0.128	-0.040	0.244	0.347	0.374	0.328
Mean absolute error	0.344	0.363	0.453	0.390	0.310	0.491	0.427	0.420	0.407
Mean squared error	0.178	0.251	0.272	0.170	0.134	0.258	0.305	0.329	0.292
Root mean squared error	0.422	0.501	0.521	0.413	0.367	0.508	0.552	0.573	0.540
Theil's U1	0.299	0.326	0.367	0.299	0.237	0.406	0.627	0.642	0.572
Theil's U2	0.597	0.708	0.737	0.505	0.449	0.622	0.781	0.811	0.764
C-statistic	-0.642	-0.497	-0.455	-0.744	-0.797	-0.612	-0.389	-0.341	-0.415

Note 5.75: The table evaluates the forecast results for the three turning points of the MSCI all properties all offices series. In this analysis, only the three best performing textual sentiment measures, based on the full corpus, have been applied. For each of the turning points, the forecast has been performed individually. All series have been estimated until three months before the occurrence of the turning point and then the next six periods have been predicted.

Different from the previous section the statistically assumed best model is able to outperform the other two methods. The second turning point period especially showed significant improvement.

Figure 5:35 - Turning point predictions MSCI all offices (all articles)

Note 5.76: The three graphs above illustrate the development of the forecast of the textual sentiment indicators during the occurrence of the turning points. The dependent variable in this analysis is the MSCI all properties all offices series.

SUMMARY

The previous analysis has shown that sentiment extracted from news articles is able to provide additional and efficient information about the market and its development. The application of machine learning algorithms in its purest form, however, has still not produced any convincing results. If the application of word lists is capable of outperforming the textual sentiment indicators from machine learning algorithms, then there remains the question as to why we should use machine learning for the extraction in the first place. The *BING* indicator has

shown good statistical results and seems to prove itself as the best indicator in the set of used methods.

5.6.1.4.2 SUB-CORPUS II: NO HOUSING

The dependent variables have not changed. Table 5:42 illustrates the descriptive statistics for the second part of the analysis. Different from the first set of indicators it can be seen that the minimum values are now less extreme, while the maximum values have increased for all eight indicators.

Table 5:42 - Summary of statistics (no housing)

Variable	Obs	Mean	Std. Dev.	Min	Max
All assets all properties (*MSCI*_change of growth rate)	158	0.297	0.459	0.000	1.000
All assets all offices (*MSCI*_change of growth rate)	158	0.272	0.446	0.000	1.000
AFINN	144	0.000	1.000	-3.355	2.475
BING	144	0.000	1.000	-3.608	2.614
NRC	144	0.000	1.000	-7.055	2.862
TM	144	0.000	1.000	-5.994	2.015
SVM (equal articles)	144	0.000	1.000	-2.392	2.014
MAXENT (equal articles)	144	0.000	1.000	-2.777	2.125
RANDOM FOREST (equal articles)	144	0.000	1.000	-7.048	2.280
MAXENT (all articles)	144	0.000	1.000	-2.504	2.826

Note 5.77: The table illustrates the summary of statistics for the probit analysis for the no housing sub-corpus.

The result of the ADF (Table 5:43) test remains unchanged. Again, the test statistics have all been higher than the critical value at the 1% level. Therefore, I do not suspect the presence of a unit root within the series.

I further determined the lag structure for the individual indicators with the help of the AIC. This time, the lag structure is slightly different to the previous analysis. While most of the lexicon approach models have lagged values in both models, half of the machine learning models (*SVM* and *MAXENT* I) enter the probit model at least for the first analysis unchanged.

Table 5:43 - Augmented Dickey-Fuller Test (no housing)

Variable	Test statistics	1% critical value	5% critical value	10% critical value	Obs.
All assets all properties (*MSCI*_change of growth rate)	-3.568	-3.491	-2.886	-2.576	157
All assets all offices (*MSCI*_change of growth rate)	-4.046	-3.491	-2.886	-2.576	157
AFINN	-7.031	-3.496	-2.887	-2.577	143
BING	-5.842	-3.496	-2.887	-2.577	143
NRC	-9.139	-3.496	-2.887	-2.577	143
TM	-9.249	-3.496	-2.887	-2.577	143
SVM (equal articles)	-7.964	-3.496	-2.887	-2.577	143
MAXENT (equal articles)	-8.390	-3.496	-2.887	-2.577	143
RANDOM FOREST (equal articles)	-9.471	-3.496	-2.887	-2.577	143
MAXENT (all articles)	-8.222	-3.496	-2.887	-2.577	143

Note 5.78: The table illustrates the results of the Augmented Dickey-Fuller Test. All test-statistics are above the critical values at a 1% level.

5.6.1.4.2.1 PROBIT MODEL RESULTS (NO HOUSING)

The first dependent variable is again the *MSCI* all properties binary growth rate. Table 5:44 illustrates the results. Different from the previous analysis, it can be seen that not all coefficients are significant. Notably, the indicators of the machine learning approach fail to remain significant; only the two *MAXENT* models show a significance at the 10% (*MAXENT* I) and 1% (*MAXENT* II) level. For those indicators which are significant, they again show a negative sign.

Comparing the values of the pseudo-R-squared, it can be seen that the *BING* (0.189) model again outperforms the other models to some extent. However, all values are below 0.2, and should, therefore, be seen as weak. Both significant machine learning models only show an R-squared value of 0.021 (*MAXENT* I) and 0.051 (*MAXENT* II). Overall, the quality of these indicators has decreased in comparison to the previous analysis.

Regarding the remaining diagnostic tests, the *BING* model shows the most satisfactory results. Notably, for the classification analysis, the remaining models fail to distribute evenly the observations into either one of the categories.

For the *Hosmer-Lemeshow* χ^2 test all models seem to pass it. However, the corresponding p-values are lower than in the previous analysis. They range between 0.109 and 0.888.

For the *ROC* curve, the area drops as low as 0.510 (*MAXENT* I), which indicates a nearly random behaviour of the indicator. For the *BING* model, the *ROC* curve presents an area of 0.773 and represents again the highest value.

To summarize: the no-housing news corpus has led to some significant changes in the indicators. It seems that the removal of housing-related articles has lowered the information quality for the overall market. The reason could be that those articles which did talk about residential topics also included CRE market information. As the analysis in section 5.6.1.4.1 has shown, a corpus consisting of all articles is more likely to provide statistically significant results.

Table 5:44 - Probit results: MSCI - all assets - all properties (no housing)

Dependent Variable MSC all assets all properties growth rate		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	Description	*AFINN*_Articles	*BING*_Articles	*NRC*_Articles	*TM*_Articles	Support Vector Machine	Maximum Entropy (1)	RANDOM FOREST	Maximum Entropy (2)
l.2z_*AFINN*_article	Standardized values for the lexicon approach with the *AFINN* lexicon	-0.591***							
		[0.129]							
l.z_*BING*_article	Standardized values for the lexicon approach with the *BING* lexicon		-0.743***						
			[0.150]						
l2.z_*NRC*_article	Standardized values for the lexicon approach with the *NRC* lexicon			-0.417***					
				[0.107]					
z_tm_article	Standardized values for the lexicon approach with the *TM* lexicon				-0.382***				
					[0.111]				
z_ceqart_*SVM*	Standardized values for the *SVM* algorithm based on the equalized training corpus with 3 categories					-0.016			
						[0.112]			
z_ceqart_max	Standardized values for the *MAXENT* algorithm based on the equalized training corpus with 3 categories						-0.212*		
							[0.113]		
l.z_ceqart_rf	Standardized values for the RF algorithm based on the equalized training corpus with 3 categories							-0.153	
								[0.107]	
l.z_callart_max	Standardized values for the *MAXENT* algorithm based on the full training corpus with 3 categories								-0.344***
									[0.119]
Constant		-0.621***	-0.633***	-0.596***	-0.586***	-0.549***	-0.560***	-0.556***	-0.579***
		[0.120]	[0.122]	[0.116]	[0.115]	[0.110]	[0.112]	[0.111]	[0.114]
Observations		144	144	144	144	144	144	144	144
Log-likelihood		-74.73	-70.46	-79.31	-80.86	-86.91	-85.13	-85.93	-82.50
LR Chi2		24.400	32.930	15.220	12.130	0.020	3.596	1.997	8.840
Number of lags		2	1	2	0	0	0	1	1
pseudo-R-squared		0.140	0.189	0.088	0.070	0.000	0.021	0.012	0.051
AIC		153.451	144.917	162.625	165.715	177.828	174.252	175.851	169.008
BIC		159.391	150.856	168.564	171.655	183.767	180.191	181.790	174.948
Correctly classified (%)		73.610	76.390	73.610	74.310	70.830	70.830	70.140	72.220
Sensitivity		26.190	35.710	14.290	14.290	0.000	2.380	0.000	11.900
Specificity		93.140	93.140	98.040	99.020	100.000	99.020	100.000	97.060
Hosmer-Lemeshow χ^2		3.640	6.210	13.080	8.320	10.760	5.140	11.630	6.170
Prob > χ^2		0.888	0.624	0.109	0.403	0.216	0.743	0.169	0.628
area under Receiver Operating Characteristic (*ROC*) curve		0.751	0.773	0.762	0.689	0.510	0.579	0.614	0.653

Standard errors in brackets; *** p<0.01, ** p<0.05, * p<0.1

Note 5.79: The table illustrates the probit results for the MSCI, all assets, all properties series. It can be seen that the textual sentiment indicators, based on the lexicon approach, remain highly significant at a 1% level. Especially, the BING measure does outperform the supervised learning measures according to the pseudo-R-squared value. From the four supervised learning measures, only the two MAXENT models remain significant at a 10% and 5% level. The test data set is the no housing sub-corpus.

Table 5:45 gives the results for the probit models that use the *MSCI* all offices series as the dependent variable. The results are similar to the previous analysis. Both the *SVM* and the *RANDOM FOREST* model fail to produce significant coefficients. While the Maximum Entropy I model is significant at the 5% level, all the remaining models are again highly significant. Further, all significant models carry the expected negative sign.

The results for McFadden's R-squared value have also been improved in comparison to the all properties analysis. Again, the *BING* model reaches the highest value with 0.237, while the remaining models are all below 0.200 and should, therefore, be seen as models with poor quality.

Similar to before, the results of the classification show that some models over-sort the observations into one of the categories. The *BING* model reached the highest classification score, with 77.780. The *NRC* and the *TM* model also produced a score of 77.780; however, they failed to sort the observations in a more reasonable way.

It is also worth mentioning that the *BING* model, as expected, reached the most significant area under the *ROC* curve with 0.812.

Table 5:45 - Probit results: MSCI - all assets - all office properties (no housing)

Dependent Variable *MSCI* all assets all office properties		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	Description	*AFINN*_Articles	*BING*_Articles	*NRC*_Articles	*TM*_Articles	Support Vector Machine	Maximum Entropy (1)	RANDOM FOREST	Maximum Entropy (2)
l2.z_*AFINN*_article	Standardized values for the lexicon approach with the *AFINN* lexicon	-0.687***							
		[0.139]							
l2.z_*BING*_article	Standardized values for the lexicon approach with the *BING* lexicon		-0.860***						
			[0.166]						
l2.z_*NRC*_article	Standardized values for the lexicon approach with the *NRC* lexicon			-0.473***					
				[0.110]					
l3.z_tm_article	Standardized values for the lexicon approach with the *TM* lexicon				-0.423***				
					[0.113]				
l.z_ceqart_*SVM*	Standardized values for the *SVM* algorithm based on the equalized training corpus with 3 categories					-0.135			
						[0.116]			
l2.z_ceqart_max	Standardized values for the *MAXENT* algorithm based on the equalized training corpus with 3 categories						-0.255**		
							[0.117]		
l.z_ceqart_rf	Standardized values for the RF algorithm based on the equalized training corpus with 3 categories							-0.145	
								[0.109]	
l2.z_callart_max	Standardized values for the *MAXENT* algorithm based on the full training corpus with 3 categories								-0.392***
									[0.124]
Constant		-0.773***	-0.795***	-0.726***	-0.707***	-0.659***	-0.673***	-0.660***	-0.699***
		[0.128]	[0.132]	[0.121]	[0.119]	[0.114]	[0.115]	[0.114]	[0.118]
Observations		144	144	144	144	144	144	144	144
Log-likelihood		-67.130	-62.58	-72.7	-74.87	-81.38	-79.62	-81.21	-76.69
LR Chi2		29.840	38.96	18.72	14.38	1.35	4.879	1.689	10.73
Number of lags		2	2	2	1	1	1	1	2
pseudo-R-squared		0.182	0.237	0.114	0.087	0.008	0.029	0.010	0.065
AIC		138.270	129.154	149.395	153.733	166.763	163.234	166.424	157.383
BIC		144.209	135.094	155.335	159.673	172.703	169.174	172.363	163.323
Correctly classified (%)		76.390	77.780	77.780	77.780	74.310	74.310	73.610	74.310
Sensitivity		27.030	35.140	16.220	16.220	100.000	2.700	0.000	8.110
Specificity		93.460	92.520	99.070	99.070	0.000	99.070	99.070	97.200
Hosmer-Lemeshow χ^2		2.980	7.060	11.97	10.660	15.950	4.530	12.460	7.780
Prob > χ^2		0.926	0.530	0.152	0.222	0.043	0.807	0.132	0.455
area under Receiver Operating Characteristic (*ROC*) curve		0.790	0.809	0.809	0.715	0.561	0.603	0.607	0.675

Standard errors in brackets; *** p<0.01, ** p<0.05, * p<0.1

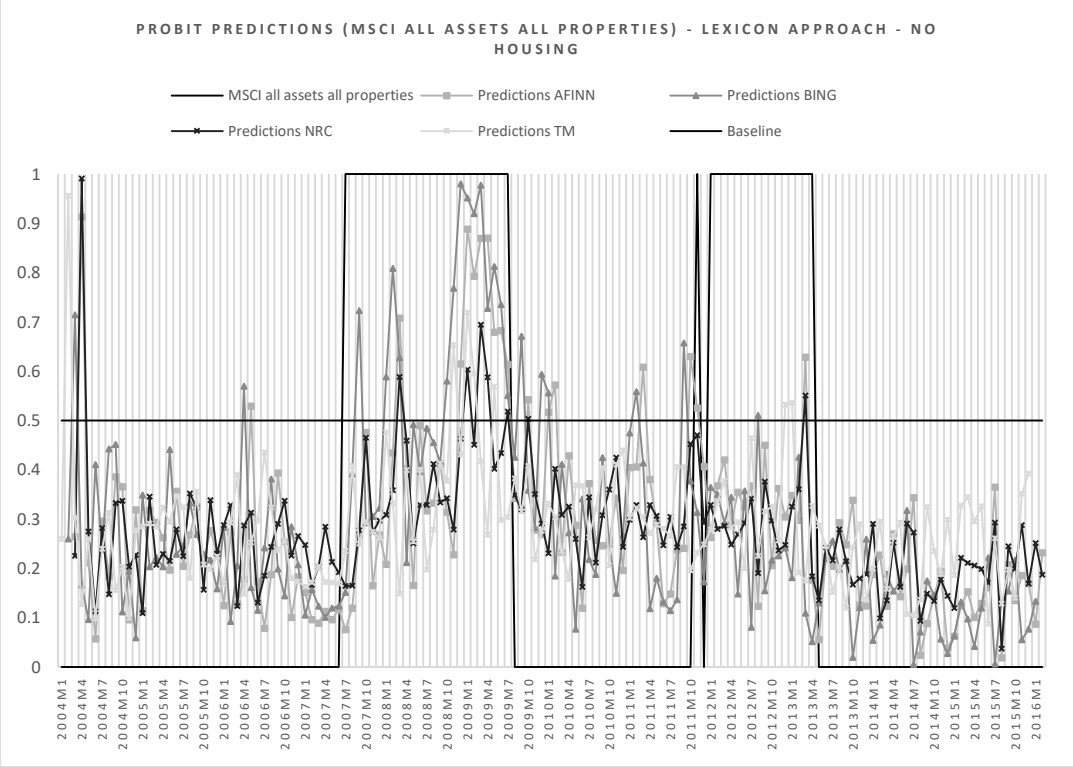
Note 5.80: The table illustrates the probit results for the MSCI, all assets, all offices series. It can be seen that the textual sentiment indicators, based on the lexicon approach, remain highly significant at a 1% level. Especially, the BING measure does outperform the supervised learning measures according to the pseudo-R-squared value. From the four supervised learning measures only the two MAXENT models are significant at a5% and 1% level. The test data set is the no housing sub-corpus.

5.6.1.4.2.2 PREDICTIONS (NO HOUSING)

Since the lexicon approach models outperform the machine learning models once more, it is not surprising to observe this superiority in the prediction graphs. While in Figure 5:36 the graphs resemble the all property dependent variable at least in the first three more substantial periods; the machine learning predictions seem more or less to fail to copy the behaviour of the *MSCI* all properties series (Figure 5:37).

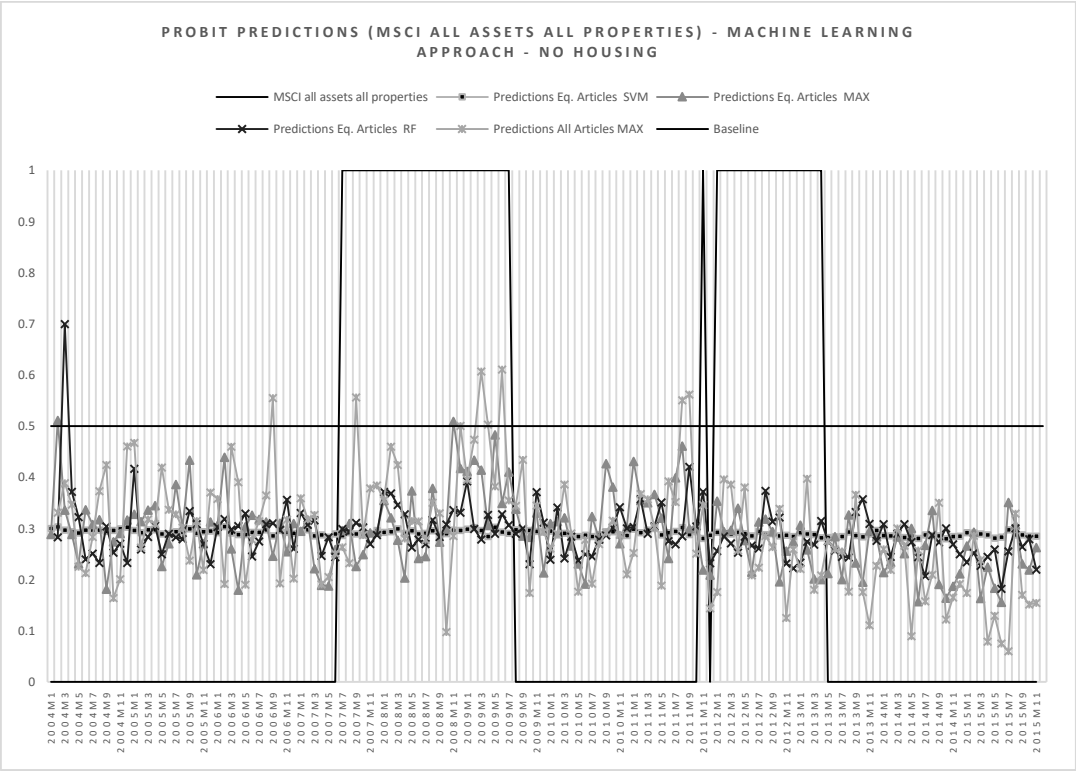
For the *BING* method, it can be seen that especially over the course of the financial crisis the probability predictions swap into the above 0.5 regions. Unfortunately, between 2011m12 and 2013m5 (negative growth) the *BING* approach did not show any amplitude towards the above 0.5 regions.

Figure 5:36 - Predictions of the MSCI all properties indicator - lexicon approach (no housing)



Note 5.81: The figure illustrates the probit predictions of the four lexicon measures, which have extracted the sentiment from the full corpus, for the MSCI all assets all properties series.

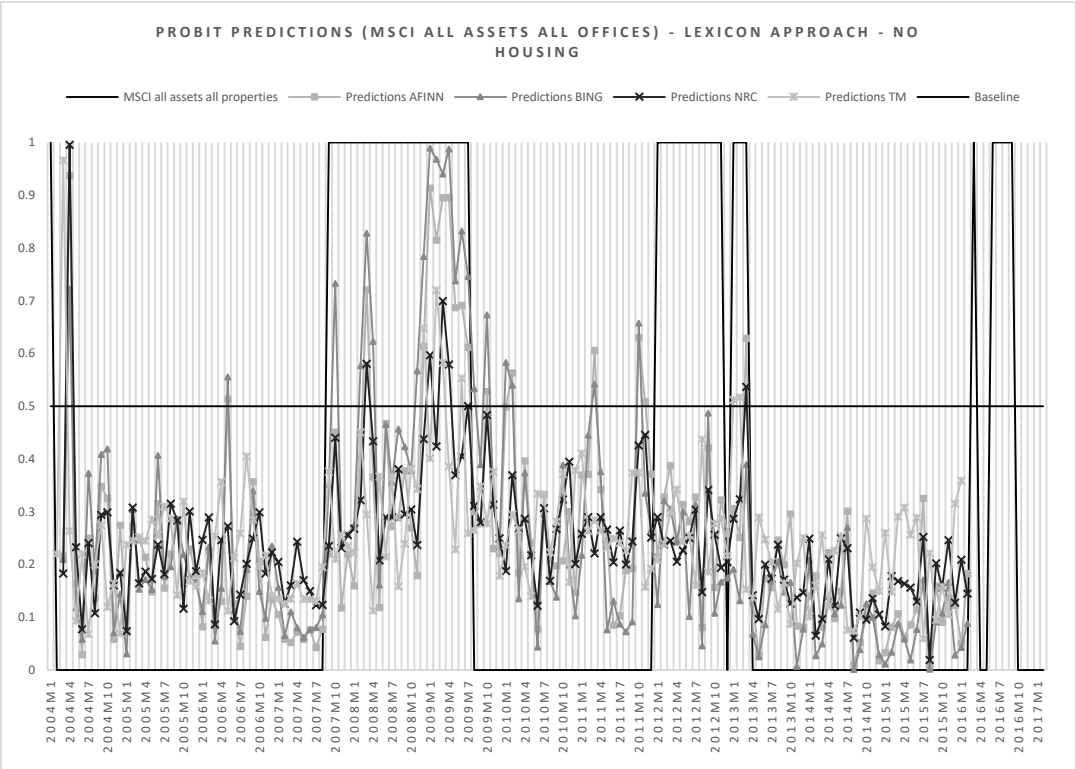
Figure 5:37 - Predictions of the MSCI all properties indicator - machine learning approach (no housing)



Note 5.82: The figure illustrates the probit predictions of the four machine learning measures, which have extracted the sentiment from the no housing sub-corpus, for the MSCI all assets all properties series.

Similar to the two probability graphs above, the superiority of the lexicon approach methods for the office dependent variable can be readily observed. Again, using the financial crisis as an example, the different indicators pick up the trend in the underlying dependent series, with the help of the extracted sentiment (Figure 5:38).

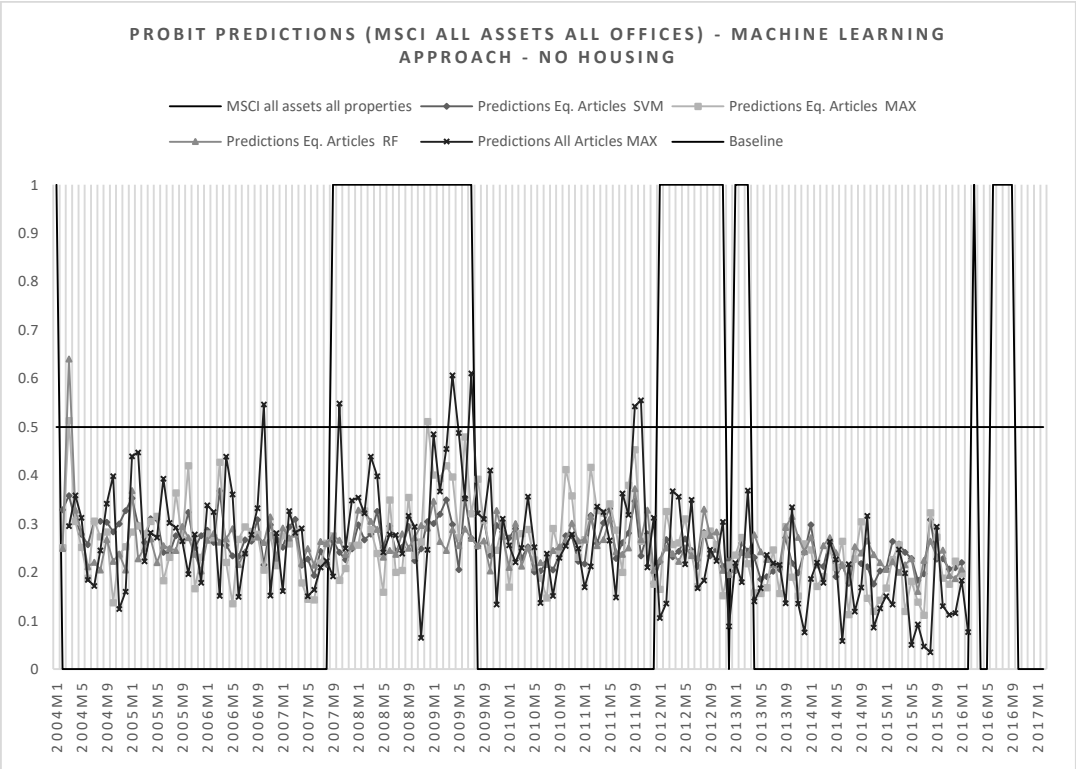
Figure 5:38 - Predictions of the MSCI all offices indicator - lexicon approach (no housing)



Note 5.83: The figure illustrates the probit predictions of the four lexicon measures, which have extracted the sentiment from the no housing sub-corpus, for the MSCI all assets all offices series.

The machine learning predictions show a much smaller resemblance to the office dependent variable. Figure 5:39 illustrates the probability results. Even though the different series show some variation, the amplitudes are less extreme and remain mostly in the below 0.5 area.

Figure 5:39 - Predictions of the MSCI all offices indicator - machine learning approach (no housing)



Note 5.84: The figure illustrates the probit predictions of the four machine learning measures, which have extracted the sentiment from the no housing sub-corpus, for the MSCI all assets all offices series.

Given these results, it appears that the machine learning methods draw most of their information from the removed articles in the underlying news corpus. The results differ remarkably to the all-articles analysis, which again proves that the lexicon approaches and here especially the *BING* method should be used to extract sentiment in a straightforward way.

5.6.1.4.2.3 DIEBOLD MARIANO TEST (NO HOUSING)

As before, I have compiled the DB test against the statistically best model. Table 5:46 and Table 5:47 illustrate the results and show that the *BING* model again outperforms the remaining models. The MSE of the *BING* model is as low as 0.160, respectively 0.140. In both cases, the *AFINN* model comes second. Different from the full corpus analysis (5.6.1.4.1.3) none of the machine learning models are capable of outperforming any of the lexicon sentiment indicators.

Table 5:46 - Diebold Mariano Test - MSCI all properties (no housing)

	MSE	Difference	S (1)	p-value
BING	0.160			
AFINN	0.172	-0.011	-0.595	0.552
NRC	0.179	-0.019	-0.937	0.349
TM	0.188	-0.027	-1.171	0.242
SVM (equal articles)	0.207	-0.047	-1.498	0.134
MAXENT (equal articles)	0.202	-0.042	-1.752	0.080
RANDOM FOREST (equal articles)	0.204	-0.044	-1.600	0.110
MAXENT (all articles)	0.194	-0.034	-1.612	0.107

Note 5.85: The table illustrates the results of the Diebold-Mariano Test for the MSCI all properties all assets series, for those indicators, which have extracted the sentiment from the no housing sub-corpus. The BING series has been used as a reference for the test and all remaining series are evaluated against it.

Table 5:47 - Diebold Mariano Test - MSCI all offices (no housing)

	MSE	Difference	S (1)	p-value
BING	0.140			
AFINN	0.152	-0.119	-0.809	0.418
NRC	0.160	-0.019	-1.178	0.238
TM	0.169	-0.027	-1.176	0.239
SVM (equal articles)	0.190	-0.048	-1.650	0.099
MAXENT (equal articles)	0.185	-0.043	-1.715	0.086
RANDOM FOREST (equal articles)	0.191	-0.049	-1.934	0.053
MAXENT (all articles)	0.175	-0.034	-1.577	0.114

Note 5.86: The table illustrates the results of the Diebold-Mariano Test for the MSCI all properties all offices series, for those indicators, which have extracted the sentiment from the no housing sub-corpus. The BING series has been used as a reference for the test and all remaining series are evaluated against it.

5.6.1.4.2.4 TURNING POINTS (NO HOUSING)

I have again chosen the three turning points for the *MSCI* all properties series. Table 5:48 shows that the two lexicon approach models have the same negative sign for the mean forecast error for the first turning point. That means that the models over predict the dependent variable. The *MAXENT* model, as well as all models for the other two turning points, do have a positive sign. Different to the previous analysis in 5.6.1.4.1.4 the forecast errors do not increase towards the third turning point.

The mean squared errors are all relatively high, with the *AFINN* model having the lowest value at 0.159 for the first turning point. This results again is surprising given the results of the DB test, where the *BING* model outperformed all remaining models.

Comparing the values of Theil's U1, only the results for the first turning point are closer to 0, rather than 1. This indicates that the models for the first turning point produce better forecasts.

The remaining forecast measures, Theil's U2 and the C-statistic show that all models outperform the naïve forecast approach. The values of Theil's U2 measure are all smaller than one, and the values of the C-statistic are negative.

Table 5:48 - Forecast evaluation for the three turning points MSCI all properties (no housing)

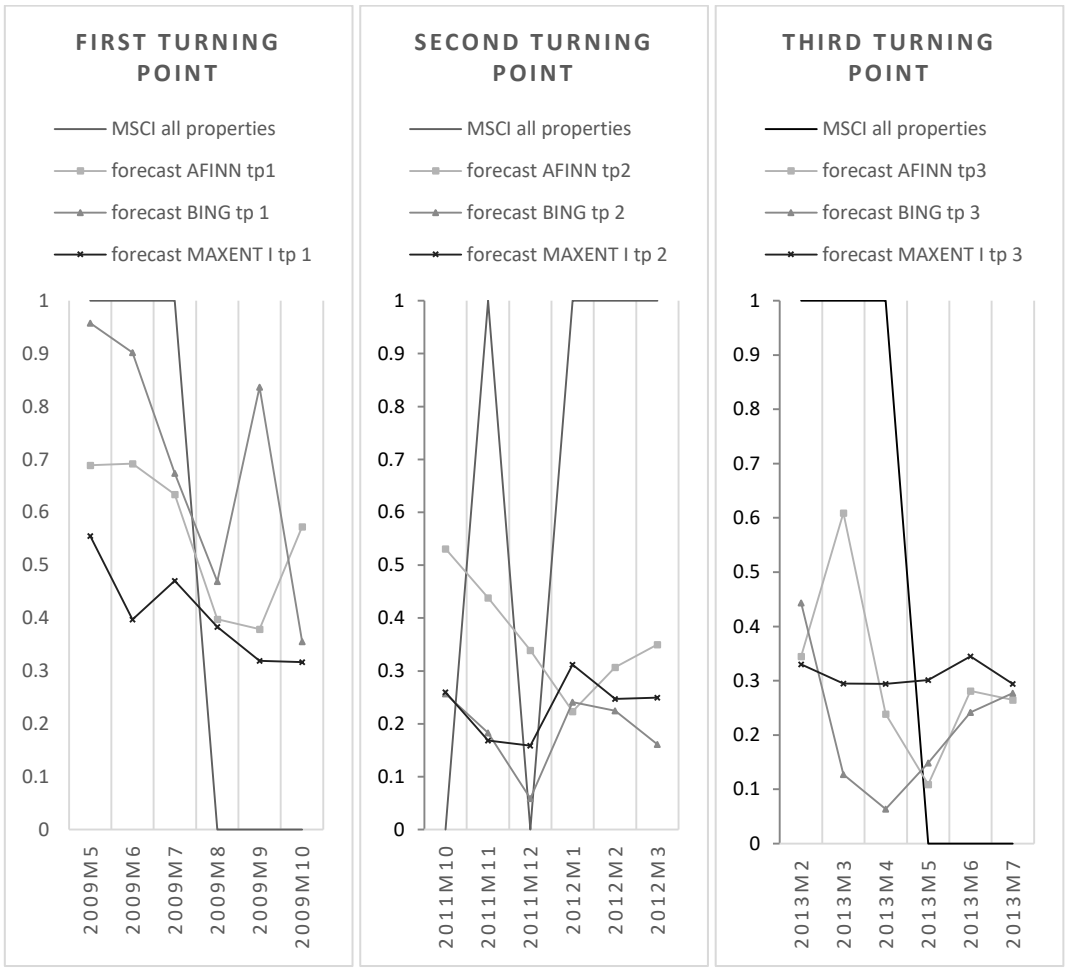
	First turning point 2009m8			Second turning point 2012m1			Third turning point 2013m5		
Measures of forecast accuracy	AFINN	BING	*MAXENT* (equal articles)	AFINN	BING	*MAXENT* (equal articles)	AFINN	BING	*MAXENT* (equal articles)
Mean forecast error	-0.060	-0.199	0.093	0.301	0.479	0.434	0.192	0.283	0.190
Mean absolute error	0.389	0.354	0.432	0.591	0.584	0.573	0.410	0.505	0.503
Mean squared error	0.159	0.194	0.198	0.369	0.436	0.398	0.220	0.351	0.290
Root mean squared error	0.399	0.440	0.445	0.608	0.660	0.630	0.469	0.592	0.538
Theil's U1	0.311	0.305	0.396	0.560	0.851	0.667	0.446	0.619	0.529
Theil's U2	0.564	0.622	0.630	0.744	0.809	0.772	0.663	0.837	0.761
C-statistic	-0.681	-0.612	-0.603	-0.445	-0.345	-0.402	-0.559	-0.298	-0.419

Note 5.87: The table evaluates the forecast results for the three turning points of the MSCI all properties all assets series. In this analysis, only the three best performing textual sentiment measures were used. For each of the turning points, the forecast has been performed individually. All series have been estimated until three months before the occurrence of the turning point and then the next six periods have been predicted.

Figure 5:40 illustrates the behaviour of the three different models over the course of the three turning points. It can be seen that, for the first turning point, all models have reacted two periods before the event takes place. Due to the occurrence of two turning points in the second

period, the behaviour of the three models is not that clear. For the last turning point, however, the *BING* model reacts again two periods ahead.

Figure 5:40 - Turning point predictions MSCI all properties (no housing)



Note 5.88: The three graphs above illustrate the development of the forecast of the textual sentiment indicators during the occurrence of the turning points. The dependent variable in this analysis is the MSCI all properties all asset series.

Looking at the all office series, the third turning point is slightly different with occurring a couple of months before the actual change sets in. Starting with the description for the forecast evaluation of the three methods, we see that the results for the mean forecast error are similar to the previous results. The *AFINN* and the *BING* model for the first turning point have a negative sign, which indicates an overreaction of the predictions. The remaining tries to show again a positive sign.

The scores of the MSE are mostly above 0.2 with the *AFINN* model once more being the best model in comparison, reaching the lowest value for the first turning point at 0.140. Compared to the previous analysis these values have improved.

Table 5:49 further reports lower Theil's U1 values for the first turning point, and increasing values for the second and third turning point, for all the models. This indicates that the models lose explanatory power over the turn of the analysis.

The last two remaining forecast measures again show that all models outperform the naïve approach. Even though the values of Theil's U2 measure are smaller than one, they are getting close to the barriers during the second turning point. The results of the C-statistics are all negative.

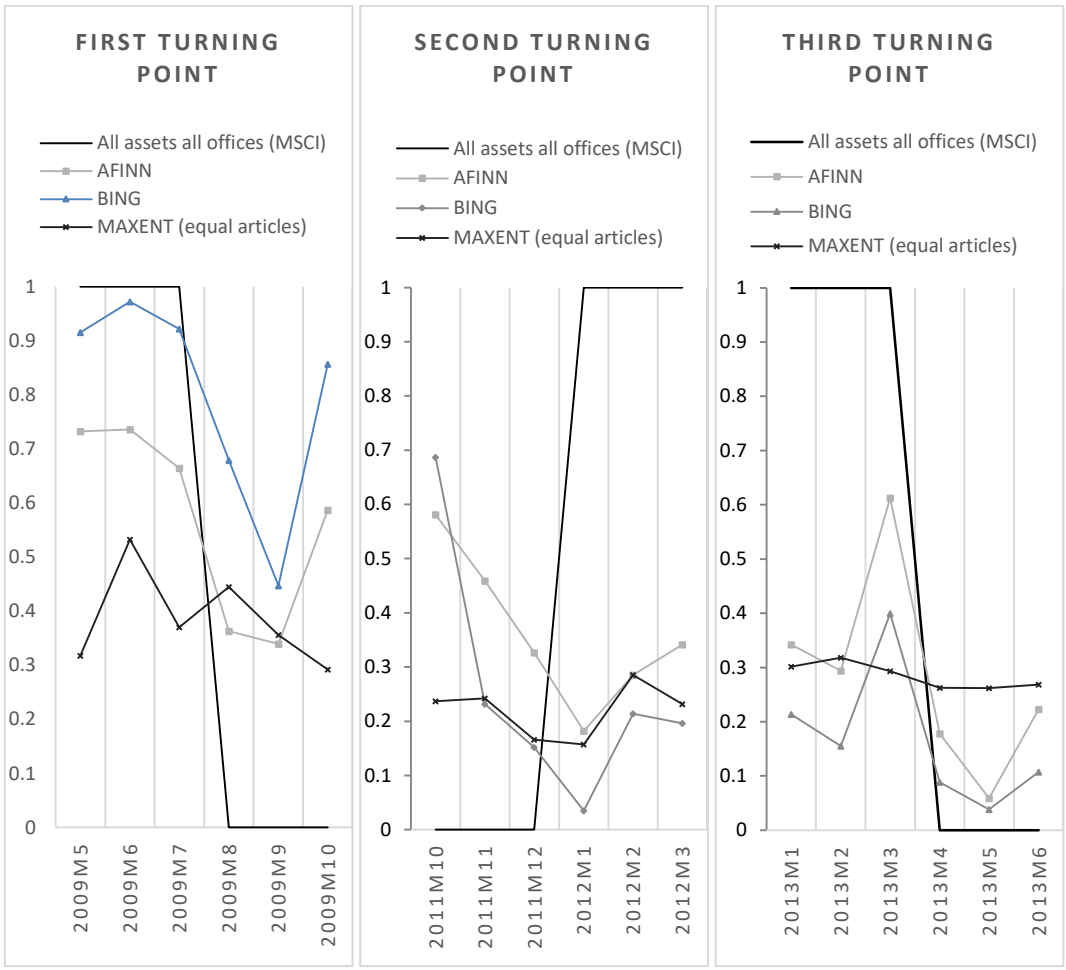
Table 5:49 - Forecast evaluation for the three turning points MSCI all offices (no housing)

	First turning point 2009m8			Second turning point 2012m1			Third turning point 2013m4		
Measures of forecast accuracy	AFINN	BING	*MAXENT* (equal articles)	AFINN	BING	*MAXENT* (equal articles)	AFINN	BING	*MAXENT* (equal articles)
Mean forecast error	-0.070	-0.299	0.114	0.137	0.247	0.280	0.215	0.333	0.215
Mean absolute error	0.359	0.362	0.478	0.592	0.604	0.495	0.368	0.411	0.480
Mean squared error	0.140	0.235	0.248	0.378	0.457	0.325	0.194	0.285	0.277
Root mean squared error	0.375	0.484	0.498	0.614	0.676	0.570	0.440	0.534	0.526
Theil's U1	0.288	0.317	0.452	0.563	0.655	0.612	0.424	0.586	0.530
Theil's U2	0.530	0.685	0.705	0.869	0.956	0.806	0.623	0.755	0.744
C-statistic	-0.718	-0.529	-0.503	-0.244	-0.085	-0.349	-0.611	-0.428	-0.445

Note 5.89: The table evaluates the forecast results for the three turning points of the MSCI all properties all offices series. In this analysis, only the three best performing textual sentiment measures were used. For each of the turning points, the forecast has been performed individually. All series have been estimated until three months before the occurrence of the turning point and then the next six periods have been predicted.

Looking at Figure 5:41, both the *AFINN* and the *BING* model react prior to the actual event of the turning point in the first period. As the forecast evaluation has shown, the results of the second forecast are less convincing. The last turning point shows more or less the right directional behaviour of the forecasts made by the *AFINN* and the *BING* approach.

Figure 5:41 - Turning point predictions MSCI all offices (no housing)



Note 5.90: The three graphs above illustrate the development of the forecast of the textual sentiment indicators during the occurrence of the turning points. The dependent variable in this analysis is the MSCI all properties all offices series.

SUMMARY

To summarize, the removal of nearly 40% of the articles has reduced the overall performance of the indicators. For both the more general and the specific office *MSCI* series, the explanatory power has dropped. Notably, the machine learning indicators have been unable to produce any convincing results. It can, therefore, be argued that the number of articles in the test corpus matters as well. More articles within a month, which discuss a similar topic, provide a better understanding of the underlying market sentiment. This observation is similar to my findings in the Natural Language Processing chapter. However, it can further be argued that articles which did contain residential terminology might carry more general information about the CRE market and should therefore not be ignored.

5.6.1.4.3 SUB-CORPUS III: LONDON

In this third analysis, I will focus on the CRE market sentiment of London. This sub-corpus includes 74,266 articles, which is slightly more than the no housing corpus. Since the capital of the U.K. represents the most extensive individual real estate market in the country, it is very likely that the sentiment towards the city is expressed in the linked articles.

Table 5:50 shows the descriptive statistics. Besides the change in minimum and maximum values, it is striking that the number of observations is much lower in comparison. As I stated earlier, the reasons are not completely clear.

Table 5:50 - Summary of statistics (London)

Variable	Obs	Mean	Std. Dev.	Min	Max
All assets all properties (*MSCI*_change of growth rate)	158	0.297	0.459	0.000	1.000
All assets all offices (*MSCI*_change of growth rate)	158	0.272	0.446	0.000	1.000
AFINN	111	0.000	1.000	-3.889	1.545
BING	111	0.000	1.000	-3.429	1.501
NRC	111	0.000	1.000	-7.770	1.355
TM	111	0.000	1.000	-7.207	1.725
SVM (equal articles)	111	0.000	1.000	-4.066	2.289
MAXENT (equal articles)	111	0.000	1.000	-5.734	1.970
RANDOM FOREST (equal articles)	111	0.000	1.000	-6.603	2.313
MAXENT (all articles)	111	0.000	1.000	-5.899	1.960

Note 5.91: The table illustrates the summary of statistics for the probit analysis for the London sub-corpus.

Similar to the two previous cases Table 5:51 does not reveal any signs of unit roots. All eight test statistics are higher than the corresponding critical value at the 1% level.

Table 5:51 - Augmented Dickey-Fuller Test (London)

Variable	Test statistics	1% critical value	5% critical value	10% critical value	Obs.
All assets all properties (*MSCI*_change of growth rate)	-3.568	-3.491	-2.886	-2.576	157
All assets all offices (*MSCI*_change of growth rate)	-4.046	-3.491	-2.886	-2.576	157
AFINN	-5.612	-3.507	-2.889	-2.579	109
BING	-4.286	-3.507	-2.889	-2.579	109
NRC	-9.088	-3.507	-2.889	-2.579	109
TM	-8.701	-3.507	-2.889	-2.579	109
SVM (equal articles)	-6.066	-3.507	-2.889	-2.579	109
MAXENT (equal articles)	-5.793	-3.507	-2.889	-2.579	109
RANDOM FOREST (equal articles)	-7.735	-3.507	-2.889	-2.579	109
MAXENT (all articles)	-6.829	-3.507	-2.889	-2.579	109

Note 5.92: The table illustrates the results of the Augmented Dickey-Fuller Test. All test-statistics are above the critical values at a 1% level.

5.6.1.4.3.1 PROBIT MODEL RESULTS (LONDON)

Since the results for the eight different indicators have not revealed any problems, they enter the two probit models. I start the description of the results again with the all properties *MSCI* converted growth rate. Table 5:52 presents the results. It can be seen that only two indicators enter the probit model with one lag (*AFINN* and *NRC*), while the remaining indicators do not have any lags.

The coefficients of the eight indicators are again negative, which is once more in line with my expectations. However, another drop in the significance of the coefficients can be observed. While the four indicators based on the lexicon approach are all significant, at least at the 5% level (*TM*), only the two *MAXENT* machine learning indicators are significant at the 5% level. The remaining two indicators fail to show any insignificance.

This result is further translated into the pseudo-R-squared value. The *BING* model is once more the best model and reaches a value of 0.14. This is again followed by the *AFINN* model (0.089). For the machine learning models, only the *MAXENT* (2) model produces a slightly higher R-squared value (0.042) than the lowest lexicon approach model (0.036).

Regarding the remaining diagnostic tests, the *BING* model shows satisfactory results. Looking at the classification analysis most of the models fail to distribute evenly the observations into either one of the two categories and overestimate one.

The results of the *Hosmer-Lemeshow* χ^2 test show that all but the *TM* model pass the test. The *TM* model only reaches a p-value of 0.024.

The last test looks at the area below the *ROC* curve. Both the *BING* and the *AFINN* model are the statistically speaking best models and merely reach a value of 0.708 and 0.707 respectively. In comparison to the other models and the previous analysis, these results are slightly lower. The *NRC* model shows the most significant value below the *ROC* curve with 0.746; however, it produced weaker results in general.

Table 5:52 - Probit results: MSCI - all assets - all properties (London)

Dependent Variable *MSCI* all assets all properties		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	Description	*AFINN*_Articles	*BING*_Articles	*NRC*_Articles	*TM*_Articles	Support Vector Machine	Maximum Entropy (1)	RANDOM FOREST	Maximum Entropy (2)
l.z_*AFINN*_article	Standardized values for the lexicon approach with the *AFINN* lexicon	-0.457***							
		[0.133]							
z_*BING*_article	Standardized values for the lexicon approach with the *BING* lexicon		-0.607***						
			[0.147]						
l1.z_*NRC*_article	Standardized values for the lexicon approach with the *NRC* lexicon			-0.289***					
				[0.111]					
z_tm_article	Standardized values for the lexicon approach with the *TM* lexicon				-0.265**				
					[0.114]				
z_ceqart_*SVM*	Standardized values for the *SVM* algorithm based on the equalized training corpus with 3 categories					-0.168			
						[0.122]			
z_ceqart_max	Standardized values for the *MAXENT* algorithm based on the equalized training corpus with 3 categories						-0.241**		
							[0.120]		
z_ceqart_rf	Standardized values for the RF algorithm based on the equalized training corpus with 3 categories							-0.103	
								[0.119]	
z_callart_max	Standardized values for the *MAXENT* algorithm based on the full training corpus with 3 categories								-0.295**
									[0.119]
Constant		-0.323**	-0.319**	-0.325***	-0.323***	-0.313**	-0.316***	-0.312**	-0.322***
		[0.126]	[0.129]	[0.123]	[0.123]	[0.122]	[0.122]	[0.121]	[0.123]
Observations		111	111	111	111	111	111	111	111
Log-likelihood		-67.11	-63.35	-70.36	-70.99	-72.67	-71.58	-73.25	-70.55
LR Chi2		13.03	20.54	6.532	5.264	1.91	4.077	0.736	6.151
Number of lags		1	0	1	0	0	0	0	0
pseudo-R-squared		0.089	0.140	0.044	0.036	0.013	0.028	0.005	0.042
AIC		138.213	130.701	144.713	145.981	149.335	147.168	150.509	145.094
BIC		143.632	136.120	150.132	151.400	154.754	152.587	155.928	150.513
Correctly classified (%)		70.270	72.070	65.770	64.860	63.960	66.670	62.160	65.770
Sensitivity		33.330	45.240	11.900	11.900	9.520	14.290	2.380	19.050
Specificity		92.750	88.410	98.550	97.100	97.100	98.550	98.550	94.200
Hosmer-Lemeshow χ^2		8.290	4.940	17.650	12.710	5.820	11.810	2.730	9.230
Prob > χ^2		0.405	0.764	0.024	0.122	0.668	0.160	0.950	0.323
area under Receiver Operating Characteristic (*ROC*) curve		0.707	0.708	0.746	0.717	0.569	0.599	0.578	0.685

Standard errors in brackets; *** p<0.01, ** p<0.05, * p<0.1

Note 5.93: The table illustrates the probit results for the MSCI, all assets, all properties series. It can be seen that all lexicon-based sentiment indicators, except for the TM model (5%), remain highly significant at a 1% level. Especially, the BING measure does outperform the supervised learning measures according to the pseudo-R-squared value. From the four supervised learning measures, only the two MAXENT models are significant at a 5% level. As a test data set the London sub-corpus was used.

Table 5:53 presents the results for the London sub-corpus with the converted *MSCI* all offices growth rate. As expected the overall performance of the various indicators increased due to the fact that the dependent variable now matches much more the extracted sentiment.

It can be seen that all but one indicator (*RANDOM FOREST*) are significant at the 5% level with the majority being significant at the 1% level. Still, the sign for all model coefficients remains negative. The increased number of highly significant coefficients is also mirrored in the pseudo-R-squared values. Model 2 once more outperforms the remaining models with a value of 0.168, followed by the *AFINN* model (0.114). The machine learning models do produce weaker results, with the two *MAXENT* models being superior in comparison.

Regarding the classification of the individual observations, most of the models underestimate the share of the sensitivity part. Only the *BING* and *AFINN* models produce reasonable results and therefore reach the highest classification scores.

Comparing the results of the *Hosmer-Lemeshow* χ^2 test, all models except the *NRC* model pass the test and show p-values above the 5% hurdle. Given that the scores for the area under the *ROC* curve are highest for the four lexicon approach models, the remaining four models only produce values below 0.7.

Table 5:53 - Probit results MSCI - all assets - all office properties (London)

Dependent Variable MSCI all offices		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	Description	AFINN_Articles	BING_Articles	NRC_Articles	TM_Articles	Support Vector Machine	Maximum Entropy (1)	RANDOM FOREST	Maximum Entropy (2)
l.z_*AFINN*_article	Standardized values for the lexicon approach with the *AFINN* lexicon	-0.512*** [0.135]							
l.z_*BING*_article	Standardized values for the lexicon approach with the *BING* lexicon		-0.658*** [0.149]						
l.z_*NRC*_article	Standardized values for the lexicon approach with the *NRC* lexicon			-0.318*** [0.111]					
z_tm_article	Standardized values for the lexicon approach with the *TM* lexicon				-0.273** [0.115]				
z_ceqart_*SVM*	Standardized values for the *SVM* algorithm based on the equalized training corpus with 3 categories					-0.262** [0.126]			
z_ceqart_max	Standardized values for the *MAXENT* algorithm based on the equalized training corpus with 3 categories						-0.296** [0.121]		
l.z_ceqart_rf	Standardized values for the RF algorithm based on the equalized training corpus with 3 categories							-0.14 [0.119]	
l.z_callart_max	Standardized values for the *MAXENT* algorithm based on the full training corpus with 3 categories								-0.313*** [0.120]
Constant		-0.460*** [0.129]	-0.463*** [0.133]	-0.455*** [0.126]	-0.424*** [0.125]	-0.417*** [0.124]	-0.420*** [0.125]	-0.435*** [0.124]	-0.450*** [0.126]
Observations		111	111	111	111	111	111	111	111
Log-likelihood		-62.63	-58.81	-66.73	-68.6	-69.09	-68.3	-69.98	-67.27
LR Chi2		16.05	23.68	7.837	5.458	4.477	6.055	1.339	6.762
Number of lags		1	1	1	0	0	0	1	1
pseudo-R-squared		0.114	0.168	0.055	0.038	0.031	0.042	0.009	0.047
AIC		129.259	121.629	137.468	138.460	142.175	140.597	143.966	138.544
BIC		134.678	127.048	142.887	143.879	147.594	146.016	149.386	143.963
Correctly classified (%)		73.87	75.68	66.67	67.57	66.67	68.47	65.77	70.27
Sensitivity		32.43	40.54	2.7	5.41	10.53	10.53	0	16.22
Specificity		94.59	93.24	98.65	98.65	95.89	98.63	98.65	97.3
Hosmer-Lemeshow χ^2		8.82	5.3	19.77	11.94	5.67	12.49	4.64	9.57
Prob > χ^2		0.357	0.724	0.011	0.154	0.684	0.130	0.794	0.296
area under Receiver Operating Characteristic (*ROC*) curve		0.729	0.735	0.770	0.737	0.619	0.639	0.598	0.696

Standard errors in brackets; *** p<0.01, ** p<0.05, * p<0.1

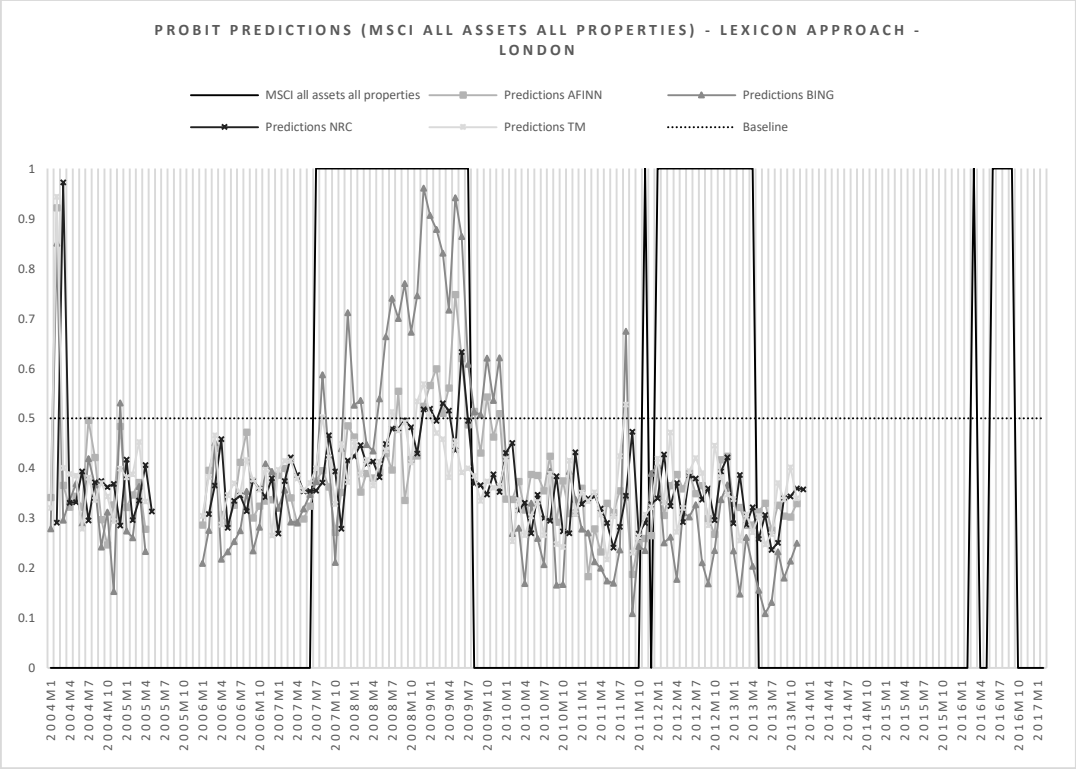
Note 5.94: The table illustrates the probit results for the MSCI, all assets, all offices series. It can be seen that three of the textual sentiment indicators (AFINN; BING and NRC), based on the lexicon approach, remain highly significant at a 1% level. Again, the SVM and the two MAXENT models show the expected negative sign and a significance at an 1% (MAXENT II), respectively 5% level (MAXENT I and SVM). Especially, the BING measure does outperform the supervised learning measures according to the pseudo-R-squared value. As a test data set the London sub-corpus was used.

To summarize, it seems that narrowing the focus of both the corpus and the dependent variable helps to produce slightly better results. Still, the produced results do not match the results based on the overall corpus. Yet, they allow us to generalize that the sentiment towards an asset class within a specific location is incorporated in the articles and can be used to anticipate the possible behaviour of the market.

5.6.1.4.3.2 PREDICTIONS (LONDON)

Figure 5:42 illustrates the predictions of the four lexicon approach models. As the above-presented analysis has shown, the *BING* model has produced the best results. However, similar to the other models *BING* also fails to pick up the negative growth between 2011m12 and 2013m5.

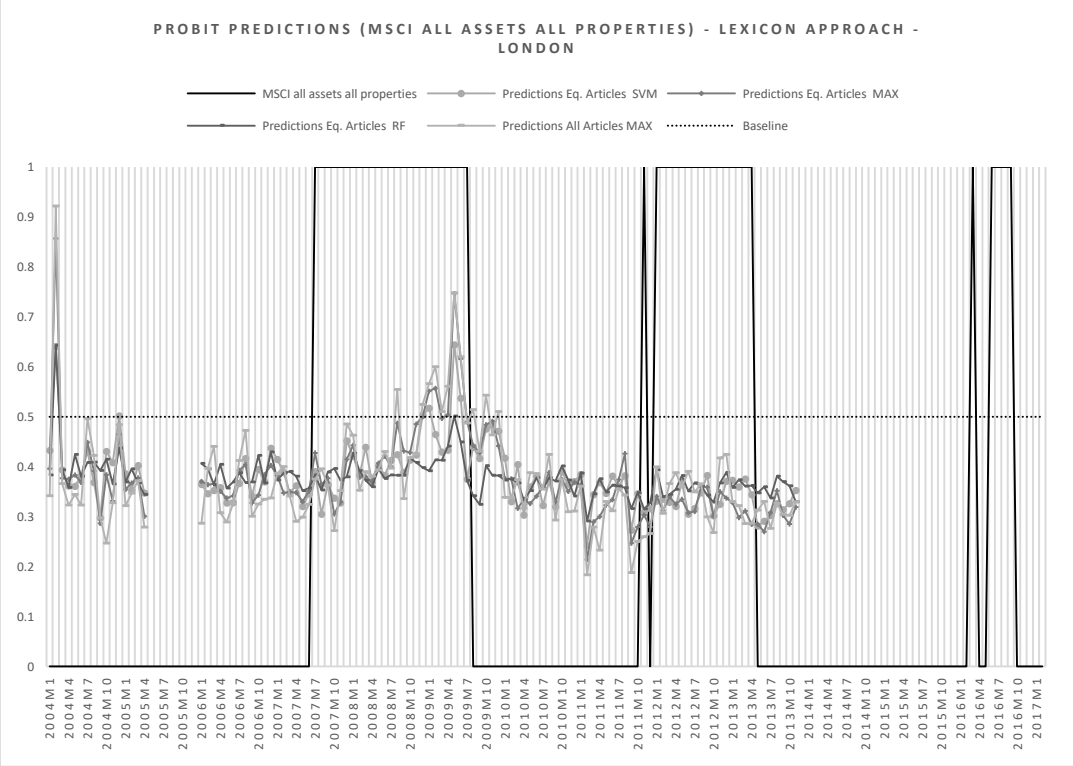
Figure 5:42 - Predictions of the MSCI all properties indicator - lexicon approach (London)



Note 5.95: The figure illustrates the probit predictions of the four lexicon-based sentiment measures, which have extracted the sentiment from the London sub-corpus, for the MSCI all assets all properties series.

The predictions of the machine learning sentiment indicators show little to no variation over the course of the analysis. There is merely a difference between positive and negative growth. Figure 5:43 summarizes the statistically weak results of the four indicators.

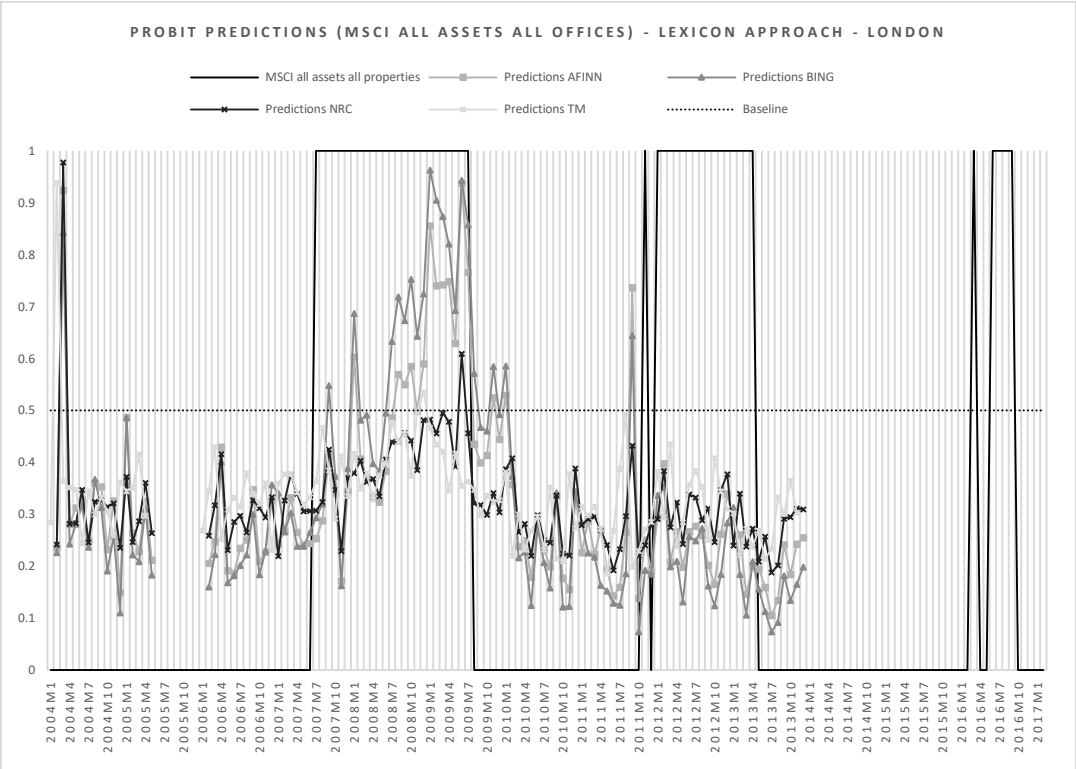
Figure 5:43 - Predictions of the MSCI all properties indicator - machine learning approach (London)



Note 5.96: The figure illustrates the probit predictions of the four machine learning measures, which have extracted the sentiment from the London sub-corpus, for the MSCI all assets all properties series.

Looking at the more distinct dependent variable, it can be seen that the results of lexicon-based sentiment indicators have improved in comparison to Figure 5:42. Next, to the *BING* model, the *AFINN* model is now also able to resemble negative growth in the period between 2007m6 and 2009m9. Yet, Figure 5:44 also shows that the indicators fail to pick up the negative growth over the period between 2011m12 and 2013m5.

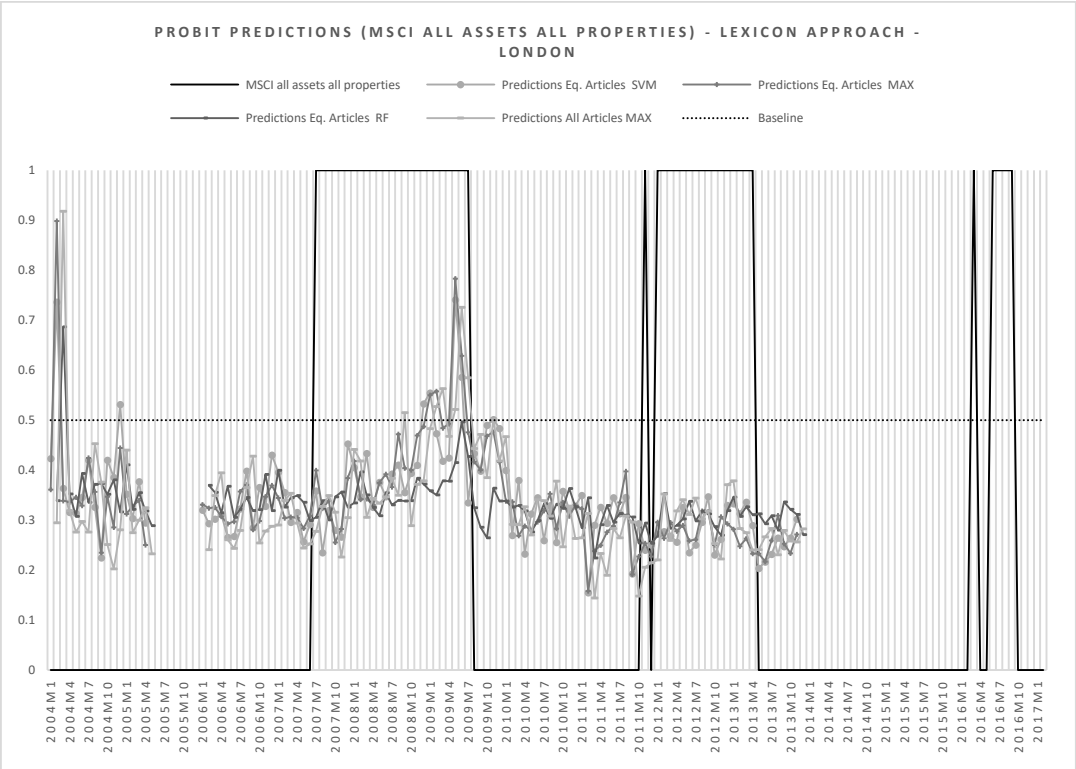
Figure 5:44 - Predictions of the MSCI all offices indicator - lexicon approach (London)



Note 5.97: The figure illustrates the probit predictions of the four lexicon measures, which have extracted the sentiment from the London sub-corpus, for the MSCI all assets all offices series.

Finally, Figure 5:45 illustrates the results of the machine learning based sentiment indicators. The change of the dependent variable has only slightly improved the results, as shown in the above analysis. However, looking in more detail at the predictions of the four models, it becomes apparent that the methods are unable to pick up both the positive and negative growth periods. Only towards the end of the financial crisis are the indicators able to reach values above the baseline.

Figure 5:45 - Predictions of the MSCI all offices indicator - machine learning approach (London)



Note 5.98: The figure illustrates the probit predictions of the four machine learning measures, which have extracted the sentiment from the London sub-corpus, for the MSCI all assets all offices series.

To summarize, changing the structure of the subcorpus has reduced the quality of the sentiment indicators and their predictive abilities. Notably, the results of the machine learning indicators seem to be quite sensitive to the number of articles within each sub-corpus. I have to admit that the analysis has produced different results than expected. The focus on articles with the word “London” has probably not extracted enough London focused sentiment.

5.6.1.4.3.3 DIEBOLD MARIANO TEST (LONDON)

The Diebold Mariano test confirms once more that the *BING* model produces the best results in comparison. Table 5:54 and Table 5:55 illustrate the results, with the *BING* model having the lowest MSE value of 0.193 and 0.176 respectively. Those values are slightly larger than the results of the previous models. The *AFINN* (0.220) model comes second for the all properties analyses; however, it is outperformed by the *MAXENT* (all articles) (0.206) and the *NRC* (0.206) approach for the all office analysis.

Table 5:54 - Diebold Mariano Test - MSCI all properties (London)

	MSE	Difference	S (1)	p-value
BING	0.193			
AFINN	0.220	-0.027	-1.040	0.298
NRC	0.217	-0.023	-0.848	0.396
TM	0.221	-0.028	-1.021	0.307
SVM (equal articles)	0.231	-0.039	-1.352	0.176
MAXENT (equal articles)	0.225	-0.033	-1.359	0.174
RANDOM FOREST (equal articles)	0.234	-0.041	-1.239	0.215
MAXENT (all articles)	0.220	-0.027	-1.040	0.298

Note 5.99: The table illustrates the results of the Diebold-Mariano Test for the MSCI all properties all properties series, for those indicators, which have extracted the sentiment from the London sub-corpus. The BING series has been used as a reference for the test and all remaining series are evaluated against it.

Table 5:55 - Diebold Mariano Test - MSCI all offices (London)

	MSE	Difference	S (1)	p-value
BING	0.176			
AFINN	0.208	-0.032	-1.186	0.235
NRC	0.202	-0.028	-1.101	0.271
TM	0.210	-0.034	-1.150	0.250
SVM (equal articles)	0.218	-0.041	-1.454	0.145
MAXENT (equal articles)	0.212	-0.036	-1.440	0.149
RANDOM FOREST (equal articles)	0.219	-0.045	-1.366	0.172
MAXENT (all articles)	0.206	-0.033	-1.204	0.228

Note 5.100: The table illustrates the results of the Diebold-Mariano Test for the MSCI all properties all offices series, for those indicators, which have extracted the sentiment from the London sub-corpus. The BING series has been used as a reference for the test and all remaining series are evaluated against it.

5.6.1.4.3.4 TURNING POINTS (LONDON)

Table 5:56 illustrates the results for the three turning points of the two lexical and the one machine learning approach. Compared to the previous analysis, the results are now a bit more mixed. The signs of the mean forecast errors are only negative for the first turning point. Throughout the remaining analysis, all forecast errors remain positive, meaning that the models under-predict the dependent variable.

Once more, the mean squared errors are all relatively high, with the *AFINN* model having the lowest value at 0.120 for the first turning point. This result is again surprising given the results of the DB test, where the *BING* model outperformed all remaining models.

Yet, the values of Theil's U1 show what has become apparent over the statistical analysis. Only the results of the first turning point are below 0.5, which indicates that the models for this turning point produce better forecasts.

All models outperform the naïve forecast approach. The results of the two remaining forecast measures reveal that the values for Theil's U2 are smaller than one and that the values of the C-statistic show negative signs.

Table 5:56 - Forecast evaluation for the three turning points - MSCI all properties (London)

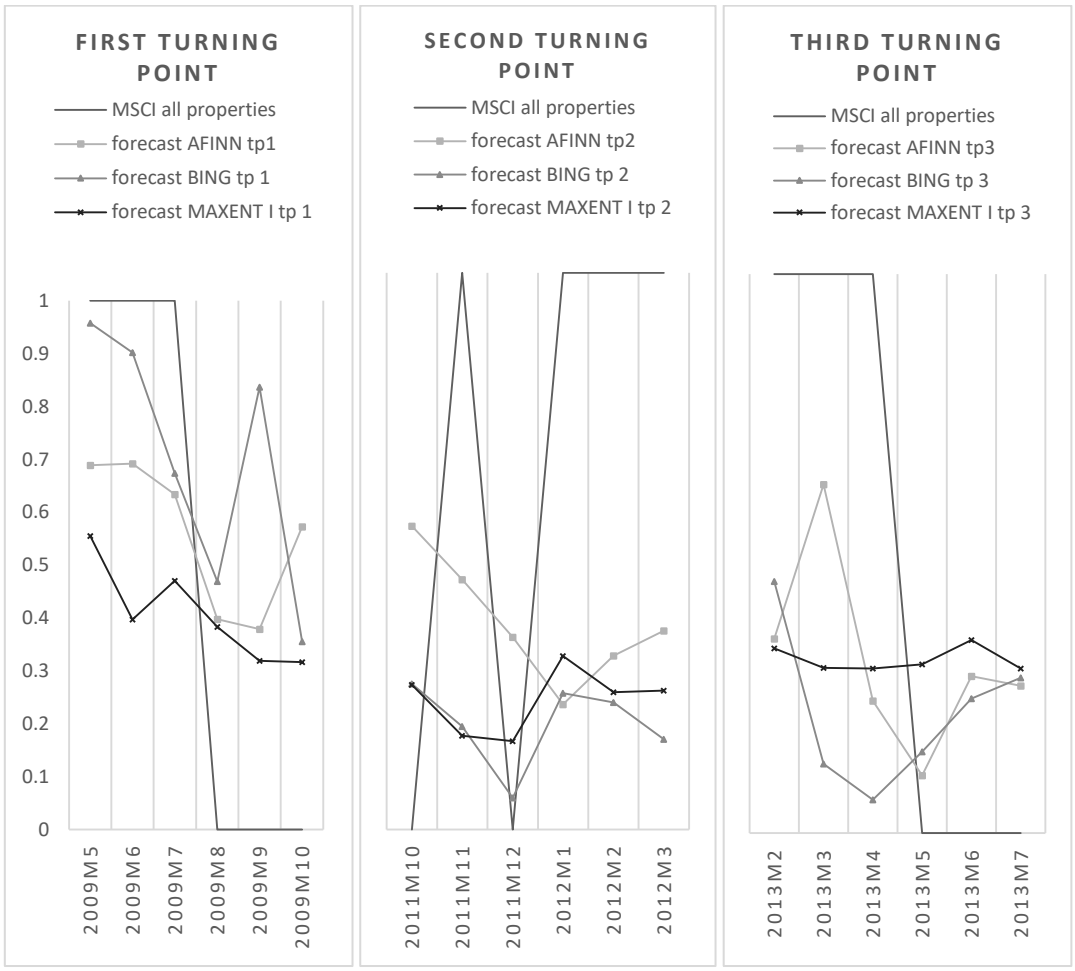
	First turning point 2009m8			Second turning point 2012m1			Third turning point 2013m5		
Measures of forecast accuracy	AFINN	BING	*MAXENT* (equal articles)	AFINN	BING	*MAXENT* (equal articles)	AFINN	BING	*MAXENT* (equal articles)
Mean forecast error	-0.158	-0.238	-0.054	0.294	0.476	0.434	0.190	0.284	0.189
Mean absolute error	0.299	0.342	0.397	0.591	0.583	0.573	0.406	0.505	0.503
Mean squared error	0.120	0.188	0.169	0.367	0.433	0.398	0.218	0.351	0.290
Root mean squared error	0.347	0.434	0.411	0.606	0.658	0.630	0.467	0.593	0.538
Theil's U1	0.248	0.295	0.322	0.554	0.845	0.667	0.443	0.619	0.529
Theil's U2	0.491	0.614	0.582	0.742	0.806	0.772	0.660	0.839	0.761
C-statistic	-0.758	-0.622	-0.661	-0.448	-0.349	-0.403	-0.563	-0.296	-0.419

Note 5.101: The table evaluates the forecast results for the three turning points of the MSCI all properties all assets series. In this analysis, only the three best performing textual sentiment measures were used. For each of the turning points, the forecast has been performed individually. All series have been estimated until three months before the occurrence of the turning point and then the next six periods have been predicted.

The graphical illustration of the predictions is given in Figure 5:46. As expected the first turning point shows the best results, with all but the *MAXENT* model reacting prior to the change of the dependent variable. For the other two turning points, only the *AFINN* and *BING* models

are able to show a consistent result by reacting more or less in accordance with the dependent variable.

Figure 5:46 - Turning point predictions, MSCI all properties (London)



Note 5.102: The three graphs above illustrate the development of the forecast of the textual sentiment indicators during the occurrence of the turning points. The dependent variable in this analysis is the MSCI all properties all assets series.

Looking at the all office series, the results for the mean forecast error are similar to the previous results. The *AFINN* and the *BING* model for the first turning point do have a negative sign, which indicates an overreaction of the predictions. The remaining methods show the opposite sign.

The scores of the MSE are generally above 0.2, except the *AFINN* model shows a value of 0.140 and 0.191 respectively for the first and third turning points. Overall these results have declined in comparison to the all properties analysis (see Table 5:57).

For the remaining measures, the results have not changed a lot. All models are still able to outperform a naïve forecast.

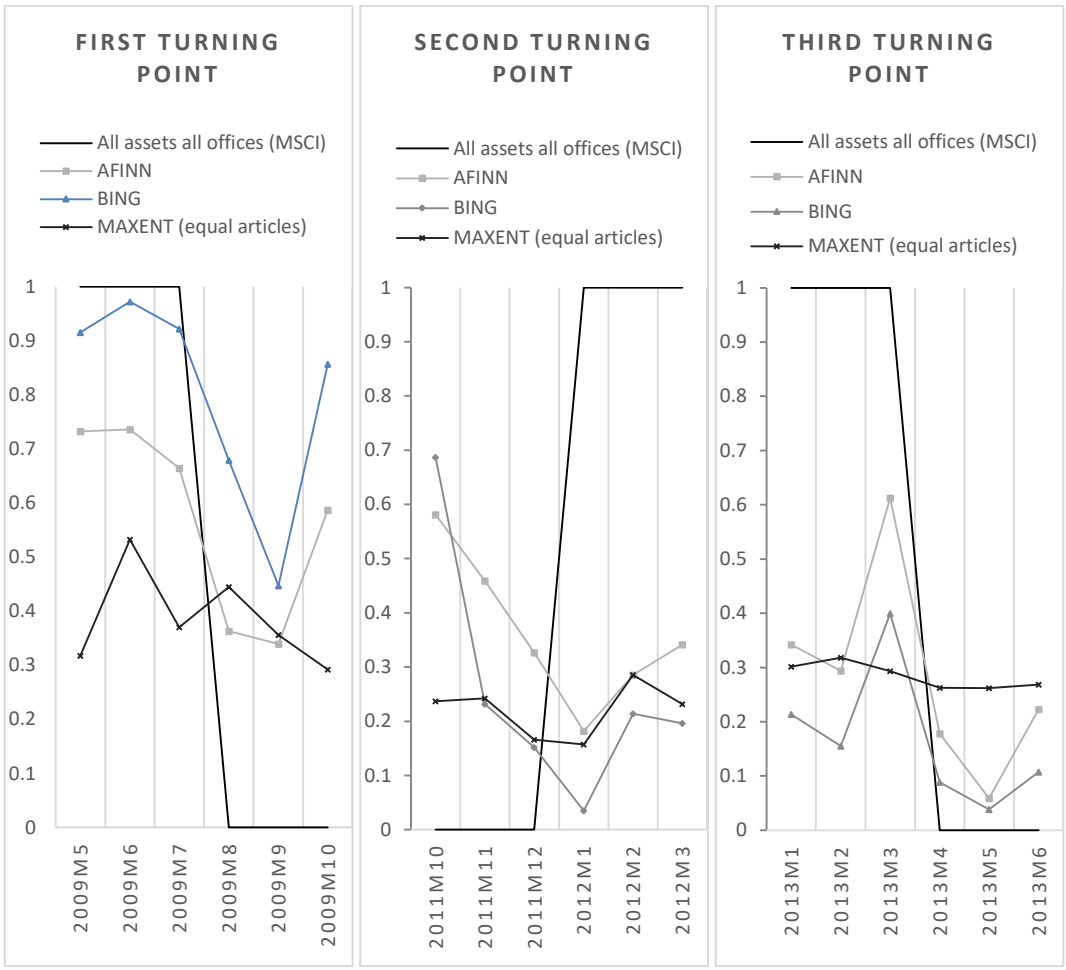
Table 5:57 - Forecast evaluation for the three turning points - MSCI all offices (London)

	First turning point 2009m8			Second turning point 2012m1			Third turning point 2013m4		
Measures of forecast accuracy	AFINN	BING	*MAXENT* (equal articles)	AFINN	BING	*MAXENT* (equal articles)	AFINN	BING	*MAXENT* (equal articles)
Mean forecast error	-0.070	-0.299	0.114	0.137	0.247	0.280	0.215	0.333	0.215
Mean absolute error	0.359	0.362	0.478	0.592	0.604	0.495	0.368	0.411	0.480
Mean squared error	0.140	0.235	0.248	0.378	0.457	0.325	0.194	0.285	0.277
Root mean squared error	0.375	0.484	0.498	0.614	0.676	0.570	0.440	0.534	0.526
Theil's U1	0.288	0.317	0.452	0.563	0.655	0.612	0.424	0.586	0.530
Theil's U2	0.530	0.685	0.705	0.869	0.956	0.806	0.623	0.755	0.744
C-statistic	-0.718	-0.529	-0.503	-0.244	-0.085	-0.349	-0.611	-0.428	-0.445

Note 5.103: The table evaluates the forecast results for the three turning points of the MSCI all properties all offices series. In this analysis, only the three best performing textual sentiment measures were used. For each of the turning points, the forecast has been performed individually. All series have been estimated until three months before the occurrence of the turning point and then the next six periods have been predicted.

Figure 5:47 illustrates the predictions of the three models over the course of the three turning points. Looking at the first graph, it can be seen that all three series react prior to the change in the dependent variable. Yet the corrections are not as extreme as expected and they do not last as long as they should. After two months, both the *AFINN* and the *BING* methods turn again towards negative growth. For the second turning point, the results are as expected and, even though the series does show some correction, they are unable to predict values of above 0.5. Finally, the last turning point shows a similar picture to the first turning point, with all series predicting the market correction two periods before the change sets in.

Figure 5:47 - Turning point predictions, MSCI all offices (London)



Note 5.104: The three graphs above illustrate the development of the forecast of the textual sentiment indicators during the occurrence of the turning points. The dependent variable in this analysis is the MSCI all properties all offices series.

SUMMARY

The analysis has revealed that the construction of the sentiment indices based on a London sub-corpus produces inferior results to the no-housing sub-corpus and especially to the overall sub-corpus. One could argue that the sentiment indicators are sensitive to the number of articles they are applied to.

5.6.1.4.4 SUB-CORPUS IV: NEWSPAPERS WITH A CIRCULATION ABOVE 100,000

The fourth sub-corpus using those articles which have been published by newspapers with a circulation of above 100,000 papers per day. The sub-corpus includes a total of 52,954 articles. The idea is that information stored in these articles reaches a wider audience and should, therefore, have a stronger impact on the real estate market.

Table 5:58 shows the descriptive statistics of the variables used in this analysis. Compared to the overall corpus the sub-corpus shows similar values for the extremes. Also, the number of observations has returned to full sample size.

Table 5:58 - Summary of statistics (100,000)

Variable	Obs	Mean	Std. Dev.	Min	Max
All assets all properties (*MSCI*_change of growth rate)	158	0.297	0.459	0.000	1.000
All assets all offices (*MSCI*_change of growth rate)	158	0.272	0.446	0.000	1.000
AFINN	144	0.000	1.000	-4.199	1.929
BING	144	0.000	1.000	-3.246	2.089
NRC	144	0.000	1.000	-8.549	2.063
TM	144	0.000	1.000	-7.304	2.130
SVM (equal articles)	144	0.000	1.000	-4.011	1.765
MAXENT (equal articles)	144	0.000	1.000	-4.572	1.677
RANDOM FOREST (equal articles)	144	0.000	1.000	-7.031	2.320
MAXENT (all articles)	144	0.000	1.000	-3.649	2.380

Note 5.105: The table illustrates the summary of statistics for the probit analysis for the 100,000 sub-corpus.

Table 5:59 illustrates the results of the Augmented Dickey-Fuller test. As before none of the eight indicators reveals any sign of a unit root.

Table 5:59 - Augmented Dickey-Fuller Test (100,000)

Variable	Test statistics	1% critical value	5% critical value	10% critical value	Obs.
All assets all properties (*MSCI*_change of growth rate)	-3.568	-3.491	-2.886	-2.576	157
All assets all offices (*MSCI*_change of growth rate)	-4.046	-3.491	-2.886	-2.576	157
AFINN	-4.532	-3.496	-2.887	-2.577	143
BING	-5.402	-3.496	-2.887	-2.577	143
NRC	-10.457	-3.496	-2.887	-2.577	143
TM	-6.970	-3.497	-2.887	-2.577	143
SVM (equal articles)	-3.642	-3.496	-2.887	-2.577	143
MAXENT (equal articles)	-5.517	-3.496	-2.887	-2.577	143
RANDOM FOREST (equal articles)	-10.348	-3.496	-2.887	-2.577	143
MAXENT (all articles)	-7.683	-3.496	-2.887	-2.577	143

Note 5.106: The table illustrates the results of the Augmented Dickey-Fuller Test. All test-statistics are above the critical values at a 1% level.

5.6.1.4.4.1 PROBIT MODEL RESULTS (100,000)

Table 5:60 shows the probit results for the all properties *MSCI* converted growth rate. The number of lags for the different indicators has been determined with the help of the AIC. The lag structure for this trial is slightly different to the previous analysis. For the lexicon-based models, the *BING* model has one lag, the *TM* model has three lags and the other two enter the probit regression without a lag. For the supervised learning indicators, only the *SVM* model has one lag.

Different to the previous analysis, all sentiment indicators are highly significant at the 1% level with the exception of the *TM* model, which is only significant at the 5% level. All constant coefficients remain highly significant.

Alongside this improvement, the pseudo-R-squared values also have improved compared to the previous two analyses. Again, the *BING* model outperforms the remaining models and reaches a value of 0.217, followed by the *AFINN* model (0.156) and the *MAXENT* I model (0.104). The supervised learning models are all better than the remaining two lexicon-based models.

For the analysis of the classification, the results are mixed. Once again, the *BING* model reaches the highest value with 78.470. Surprisingly the *AFINN* (75.00) and the two *MAXENT*

models (74.31 and 76.39) also seem to be able to sort the observations more or less appropriately.

Looking at the Hosmer-Lemeshow chi 2 test, it can be seen that most models have passed it, with p-values above 0.05. Only the *NRC* model failed the test.

For the area under the *ROC* curve, the *BING* (0.785) model outperforms the remaining models. The *AFINN* model (0.782) ranks second. However, all models except for the *TM*, the *SVM* and the *MAXENT* II model have values above 0.7.

To conclude, different from the previous results, this sub-corpus has not suffered any information loss from the reduction of the number of articles. Once more the *BING* model outperformed the remaining seven models invariably. Overall, the quality of the indicators for the all properties *MSCI* adjusted growth rate has improved in comparison to the no-housing or the London sub-corpus.

Table 5:60 - Probit results: MSCI - all assets - all properties (100,000)

Dependent Variable *MSCI* all properties		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	Description	*AFINN*_Articles	*BING*_Articles	*NRC*_Articles	*TM*_Articles	Support Vector Machine	Maximum Entropy (1)	RANDOM FOREST	Maximum Entropy (2)
z_*AFINN*_article = L,	Standardized values for the lexicon approach with the *AFINN* lexicon	-0.614***							
		[0.127]							
z_*BING*_article = L,	Standardized values for the lexicon approach with the *BING* lexicon		-0.769***						
			[0.143]						
z_*NRC*_article = L,	Standardized values for the lexicon approach with the *NRC* lexicon			-0.298***					
				[0.100]					
z_tm_article = L,	Standardized values for the lexicon approach with the *TM* lexicon				-0.253**				
					[0.105]				
z_ceqart_*SVM* = L,	Standardized values for the *SVM* algorithm based on the equalized training corpus with 3 categories					-0.344***			
						[0.117]			
z_ceqart_max	Standardized values for the *MAXENT* algorithm based on the equalized training corpus with 3 categories						-0.490***		
							[0.122]		
z_ceqart_rf	Standardized values for the RF algorithm based on the equalized training corpus with 3 categories							-0.315***	
								[0.106]	
z_callart_max	Standardized values for the *MAXENT* algorithm based on the full training corpus with 3 categories								-0.441***
									[0.120]
Constant		-0.620***	-0.633***	-0.576***	-0.546***	-0.577***	-0.593***	-0.576***	-0.593***
		[0.120]	[0.124]	[0.113]	[0.112]	[0.114]	[0.116]	[0.114]	[0.116]
Observations		144	144	144	144	144	144	144	144
Log-likelihood		-73.35	-68.07	-82.68	-84.96	-82.36	-77.87	-82.55	-79.57
LR Chi2		27.15	37.700	8.49	5.675	9.137	18.11	8.743	14.7
Number of lags		1	0	1	3	1	0	0	0
pseudo-R-squared		0.156	0.217	0.048	0.032	0.052	0.104	0.050	0.084
AIC		150.699	140.149	169.358	173.914	168.711	159.737	169.105	163.145
BIC		156.639	146.089	175.297	179.853	174.650	165.677	175.045	169.084
Correctly classified (%)		75.000	78.470	70.830	70.140	70.830	74.310	70.830	76.390
Sensitivity		26.190	42.860	2.380	2.330	9.520	26.160	4.760	28.570
Specificity		95.100	93.140	99.020	99.010	96.080	94.120	98.040	96.080
Hosmer-Lemeshow χ^2		7.250	7.260	19.330	14.680	2.990	8.620	8.830	6.640
Prob > χ^2		0.509	0.509	0.013	0.065	0.934	0.375	0.357	0.575
area under Receiver Operating Characteristic (*ROC*) curve		0.782	0.785	0.770	0.654	0.656	0.722	0.713	0.690

Standard errors in brackets; *** p<0.01, ** p<0.05, * p<0.1

Note 5.107: The table illustrates the probit results for the MSCI, all assets, all properties series. It can be seen that all textual sentiment indicators remain highly significant at a 1% level, with the exception of the TM induced model (5%). The BING measure does outperform the supervised learning measures according to the pseudo-R-squared value. As a test data set the 100,000 sub-corpus was used.

Table 5:61 illustrates the result of the all office *MSCI* modified growth rate. The number of lags remained unchanged, compared to the previous analysis. The change of the dependent variable has caused an improvement in the significance of the various indicators. All sentiment coefficients remained highly significant at the 1% level, and the *TM* model reached a significance of 5%. The coefficients of the constants for the eight different models remain highly significant at the 1% level.

Looking at the pseudo-R-squared value, it can be seen that the values have been slightly improved upon the previous try. Again, the *BING* model performs best, with a pseudo-R-squared value of 0.239; second comes the *AFINN* model with 0.186, and the *MAXENT* I ranks third with 0.139.

For the classification, the *BING* model reaches the highest value with 79.86, while the remaining models score slightly lower. It seems that again only the *NRC*, the *TM* and the Random Forrest model are unable to classify the observations appropriately.

Regarding the *Hosmer-Lemeshow* χ^2 test all but the *NRC* model pass the test. While the *BING* model has produced once more the best results, it is surprising that the *AFINN* model covers a slightly larger area under the *ROC* curve in comparison. The *BING* model reaches a value of 0.805 and the *AFINN* model a value of 0.809.

To summarize, the focus on the office market has improved the results throughout this analysis. Overall the results are better than in the previous two parts. Therefore, my above-stated argument, that the number of articles might influence the performance of the indicators cannot be entirely true.

Table 5:61 - Probit results: MSCI - all assets - all offices (100,000)

Dependent Variable *MSCI* all offices		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	Description	*AFINN*_Articles	*BING*_Articles	*NRC*_Articles	*TM*_Articles	Support Vector Machine	Maximum Entropy (1)	RANDOM FOREST	Maximum Entropy (2)
z_*AFINN*_article = L,	Standardized values for the lexicon approach with the *AFINN* lexicon	-0.664***							
		[0.130]							
z_*BING*_article = L,	Standardized values for the lexicon approach with the *BING* lexicon		-0.794***						
			[0.144]						
z_*NRC*_article = L,	Standardized values for the lexicon approach with the *NRC* lexicon			-0.322***					
				[0.101]					
z_tm_article = L,	Standardized values for the lexicon approach with the *TM* lexicon				-0.240**				
					[0.107]				
z_ceqart_*SVM* = L,	Standardized values for the *SVM* algorithm based on the equalized training corpus with 3 categories					-0.454***			
						[0.123]			
z_ceqart_max	Standardized values for the *MAXENT* algorithm based on the equalized training corpus with 3 categories						-0.574***		
							[0.129]		
z_ceqart_rf = L,	Standardized values for the RF algorithm based on the equalized training corpus with 3 categories							-0.315***	
								[0.107]	
z_callart_max	Standardized values for the *MAXENT* algorithm based on the full training corpus with 3 categories								-0.495***
									[0.124]
Constant		-0.756***	-0.777***	-0.690***	-0.673***	-0.687***	-0.705***	-0.686***	-0.696***
		[0.126]	[0.130]	[0.117]	[0.115]	[0.119]	[0.122]	[0.117]	[0.120]
Observations		144	144	144	144	144	144	144	144
Log-likelihood		-66.800	-62.410	-77.210	-79.630	-75.61	-71.570	-77.800	-74.310
LR Chi2		30.520	39.300	9.698	4.849	14.97	23.060	8.518	17.580
Number of lags		1	1	2	2	0	0	1	0
pseudo-R-squared		0.186	0.239	0.059	0.029	0.090	0.139	0.051	0.106
AIC		137.591	128.816	158.415	163.264	155.228	147.139	159.595	152.624
BIC		143.531	134.756	164.355	169.203	161.168	153.078	165.534	158.563
Correctly classified (%)		78.470	79.860	73.610	73.610	71.530	77.080	74.310	77.080
Sensitivity		27.030	40.540	0.000	0.000	10.530	28.950	2.700	23.680
Specificity		96.260	93.460	99.070	99.070	93.400	94.340	99.070	96.230
Hosmer-Lemeshow χ^2		11.340	7.550	21.020	9.330	2.850	8.420	9.030	7.200
Prob > χ^2		0.183	0.479	0.007	0.315	0.943	0.393	0.339	0.515
area under Receiver Operating Characteristic (*ROC*) curve		0.809	0.805	0.802	0.666	0.704	0.759	0.707	0.715

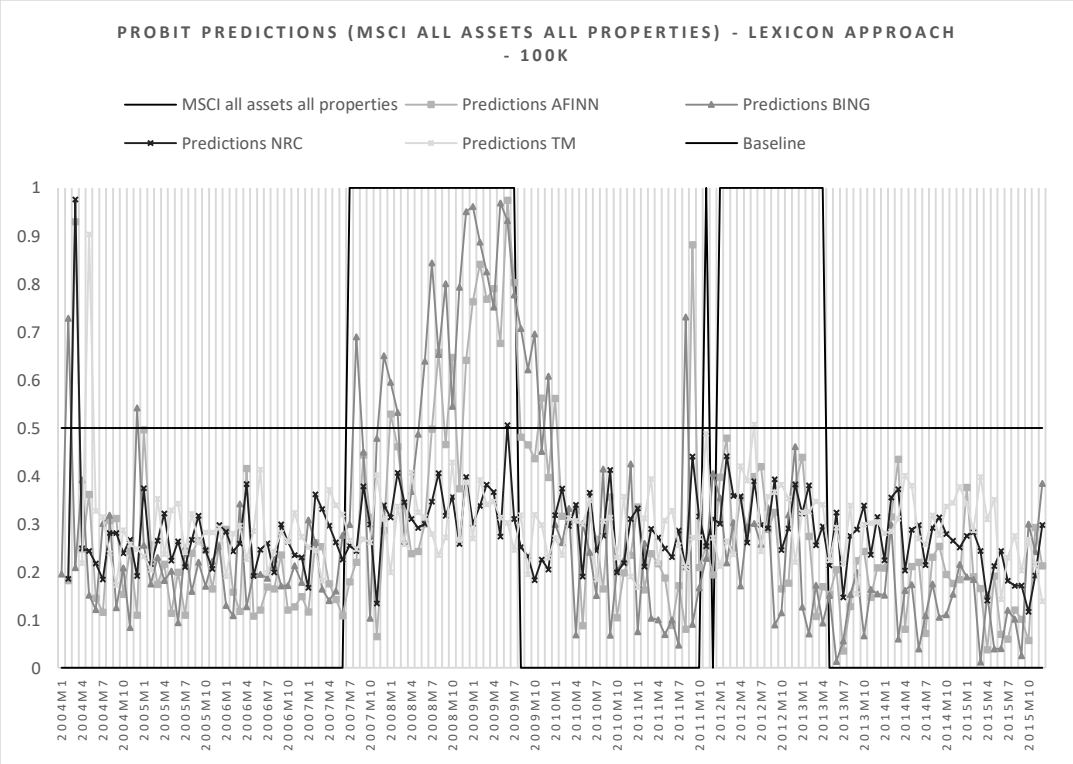
Standard errors in brackets; *** p<0.01, ** p<0.05, * p<0.1

Note 5.108: The table illustrates the probit results for the MSCI, all assets, all offices series. It can be seen that all textual sentiment indicators remain highly significant at a 1% level, with the exception of the TM induced model (5%). The BING measure does outperform the remaining measures according to the pseudo-R-squared value. As a test data set the 100,000 sub-corpus was used.

5.6.1.4.4.2 PREDICTIONS (100,000)

Figure 5:48 illustrates the prediction of the four lexicon-based indicators for the all properties series. While the indicators are able to mirror the development in times of positive growth, they fail to copy these developments in the period of negative growth. In the first period with negative growth, only the *BING* and *AFINN* models pick up the trend and follow the market movement. However, in succeeding periods, these two are also unable to react in line with the market. For the negative growth observation in 2011m11, both indicators react two to three periods prior to that event. In the more extended period of negative growth starting from 2012m2, they, unfortunately, fail to match the market.

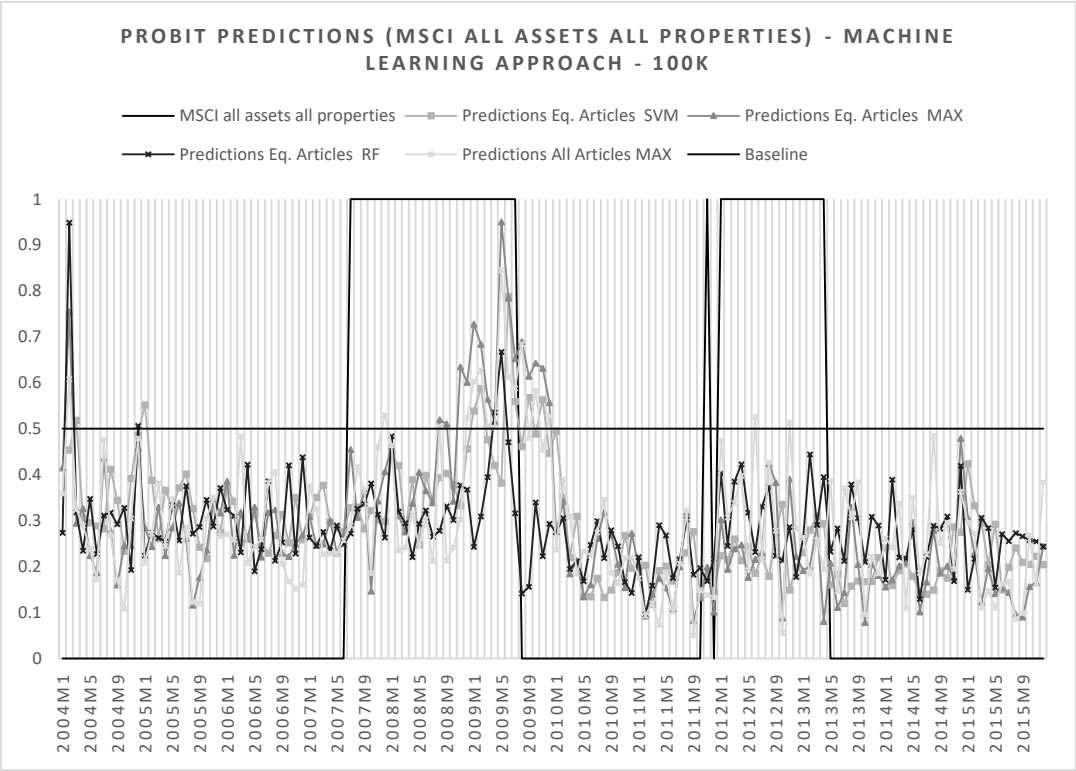
Figure 5:48 - Predictions of the MSCI all properties indicator - Lexicon approach (100,000)



Note 5.109: The figure illustrates the probit predictions of the four lexicon-based sentiment measures, which have extracted the sentiment from the 100,00 sub-corpus, for the MSCI all assets all properties series.

Looking at the machine learning algorithms (Figure 5:49) it can be seen that both *MAXENT* models, and to some extent the Random Forrest indicator, follow the market at least during the first period with negative growth. In the subsequent month, however, none of the four indicators is able to mirror the market movement.

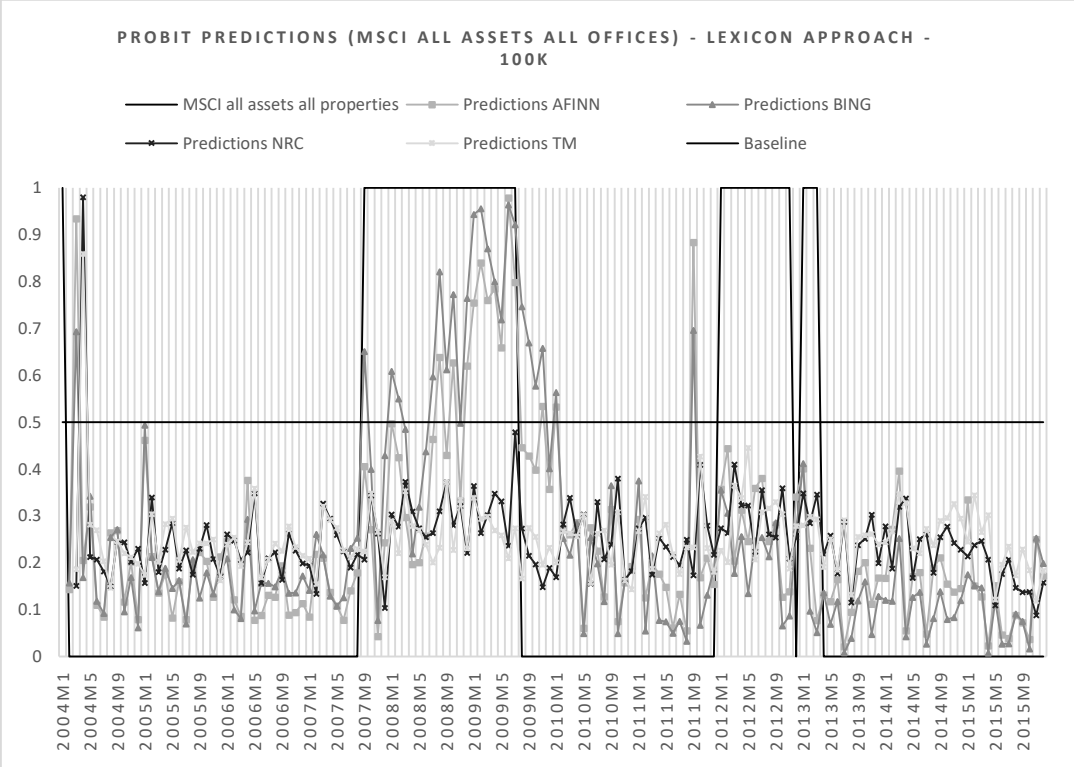
Figure 5:49 - Predictions of the MSCI all properties indicator - Machine learning approach (100,000)



Note 5.110: The figure illustrates the probit predictions of the four machine learning sentiment measures, which have extracted the sentiment from the 100,000 sub-corpus, for the MSCI all assets all properties series.

The following two graphs show the results of the all office series. The result of the lexicon approach has not changed dramatically (Figure 5:50). The *BING* and the *AFINN* model are the only two which show some market resemblance.

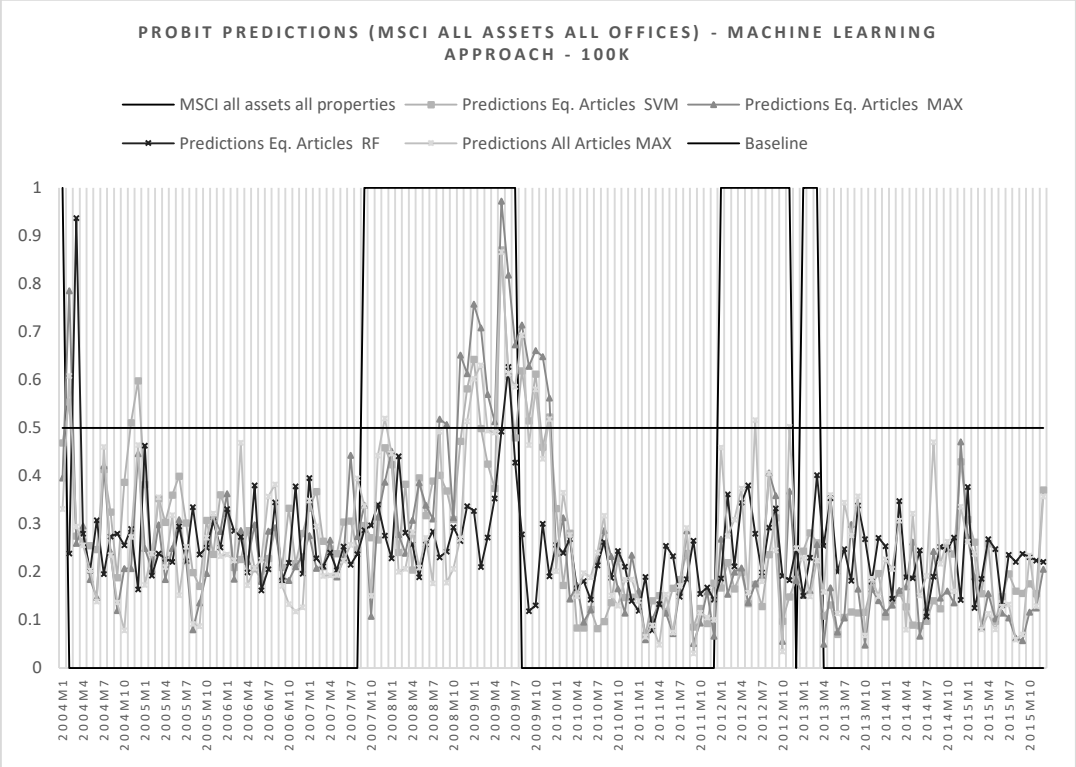
Figure 5:50 - Predictions of the MSCI all offices indicator - Lexicon approach (100,000)



Note 5.111: The figure illustrates the probit predictions of the four lexicon-based sentiment measures, which have extracted the sentiment from the 100,000 sub-corpus, for the MSCI all assets all properties series.

The same three models for the machine learning algorithms (Figure 5:51) are able to pick up the market development at least to some degree towards the end of the first period with negative growth.

Figure 5:51 - Predictions of the MSCI all offices indicator - Machine learning approach (100,000)



Note 5.112: The figure illustrates the probit predictions of the four machine learning measures, which have extracted the sentiment from the 100,000 sub-corpus, for the MSCI all assets all offices series.

5.6.1.4.4.3 DIEBOLD MARIANO TEST (100,000)

The results of the Diebold Mariano test reveal that the *BING* model still outperforms the other models. This is in-line with my expectations, given the superior results of the model in the previously described analysis. Table 5:62 shows the results for the *MSCI* all properties adjusted growth rate. The MSE value of the *BING* is as low as 0.155 and is followed by the *AFINN* (0.166) and the *MAXENT* I model (0.179). For the *MSCI* all offices models (Table 5:63), *BING* reaches a smaller value of 0.136, which again is followed by the *AFINN* model (0.147) and the *MAXENT* I approach (0.167).

Table 5:62 - Diebold Mariano Test - MSCI all properties (100,000)

	MSE	Difference	S (1)	p-value
BING	0.153			
AFINN	1.650	-0.012	-0.645	0.519
NRC	0.192	-0.039	-1.181	0.238
TM	0.201	-0.050	-1.364	0.173
SVM (equal articles)	0.195	-0.042	-1.882	0.060
MAXENT (equal articles)	0.179	-0.027	-1.750	0.080
RANDOM FOREST (equal articles)	0.192	-0.040	-1.268	0.205
MAXENT (all articles)	0.184	-0.325	-1.224	0.221

Note 5.113: The table illustrates the results of the Diebold-Mariano Test for the MSCI all properties all properties series, for those indicators, which have extracted the sentiment from the 100,000 sub-corpus. The BING series has been used as a reference for the test and all remaining series are evaluated against it.

Table 5:63 - Diebold Mariano Test - MSCI all offices (100,000)

	MSE	Difference	S (1)	p-value
BING	0.136			
AFINN	0.147	-0.010	-0.690	0.489
NRC	0.175	-0.038	-1.172	0.241
TM	0.185	-0.048	-1.370	0.170
SVM (equal articles)	0.173	-0.037	-2.109	0.034
MAXENT (equal articles)	0.161	-0.024	-1.781	0.074
RANDOM FOREST (equal articles)	0.178	-0.042	-1.344	0.178
MAXENT (all articles)	0.166	-0.030	-1.185	0.235

Note 5.114: The table illustrates the results of the Diebold-Mariano Test for the MSCI all properties all offices series, for those indicators, which have extracted the sentiment from the 100,000 sub-corpus. The BING series has been used as a reference for the test and all remaining series are evaluated against it.

5.6.1.4.4.4 TURNING POINTS (100,000)

The in-detail analysis of the three turning points for both of the *MSCI* series reveals that the *BING* model once more is not capable of dominating the other two models for these specific observations. Starting with the all properties series, Table 5:64 illustrates the forecast evaluation for the three models at the three turning points. It can be seen that the models only have a negative sign for the first turning point. For all remaining instances, the signs are positive, meaning that the models underpredict the market development.

Considering the mean squared error, it can be seen that the *BING* model is outperformed by the other two models overall at the first and second turning points. During the last period, the *BING* model ranks second after the *AFINN* model. The values for Theil's U1 are smallest for the first turning points and increase afterwards.

Checking whether the models outperform a naïve forecast, it can be seen that Theil's U2 and the C-statistic are below 1 and below 0 respectively for all instances.

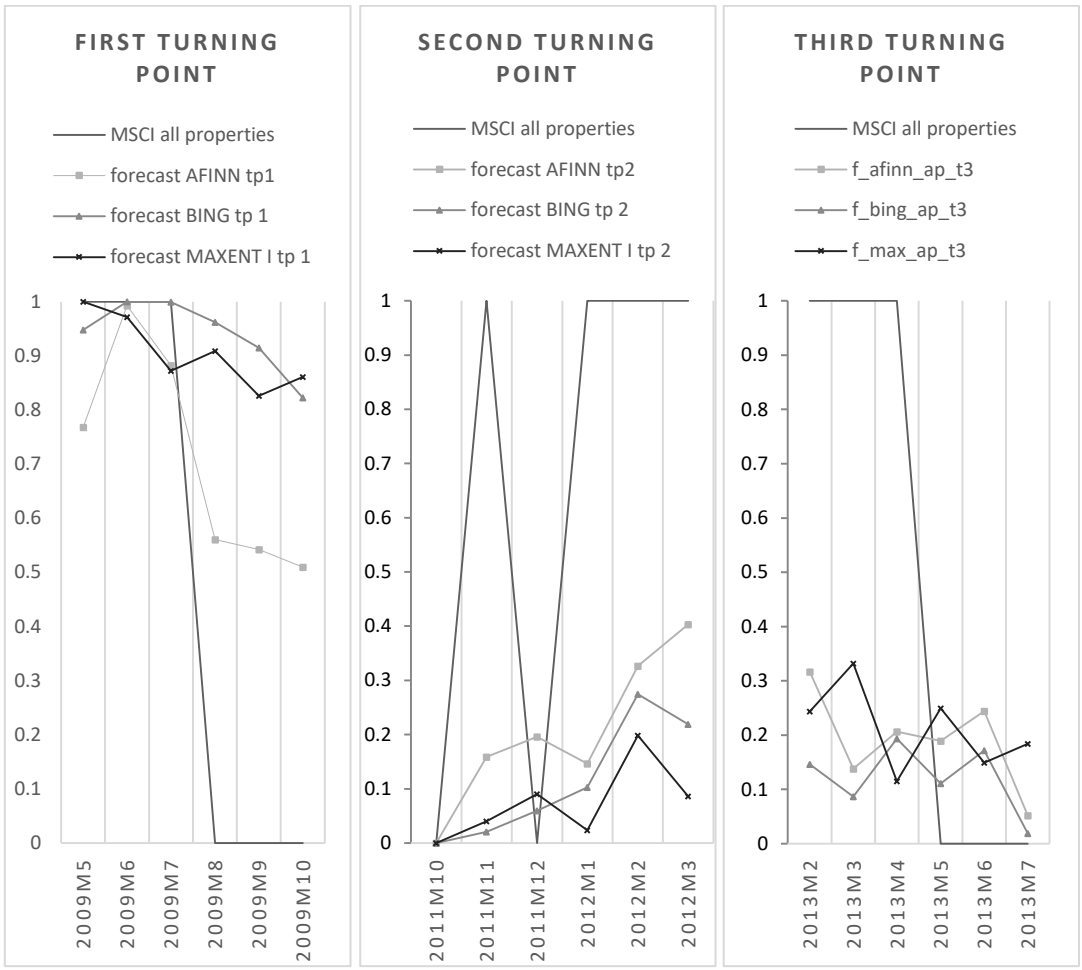
Table 5:64 - Forecast evaluation for the three turning points - MSCI all properties (100,000)

	First turning point 2009m8			Second turning point 2012m1			Third turning point 2013m5		
Measures of forecast accuracy	AFINN	BING	*MAXENT* (equal articles)	AFINN	BING	*MAXENT* (equal articles)	AFINN	BING	*MAXENT* (equal articles)
Mean forecast error	-0.209	-0.440	-0.406	0.429	0.538	0.572	0.309	0.378	0.287
Mean absolute error	0.328	0.458	0.458	0.531	0.579	0.593	0.470	0.479	0.482
Mean squared error	0.156	0.406	0.377	0.359	0.474	0.511	0.322	0.375	0.320
Root mean squared error	0.395	0.637	0.614	0.599	0.689	0.715	0.568	0.613	0.565
Theil's U1	0.274	0.386	0.380	0.623	0.938	0.874	0.620	0.728	0.607
Theil's U2	0.558	0.902	0.869	0.734	0.844	0.876	0.803	0.867	0.800
C-statistic	-0.688	-0.186	-0.244	-0.460	-0.287	-0.232	-0.354	-0.248	-0.359

Note 5.115: The table evaluates the forecast results for the three turning points of the MSCI all properties all assets series. In this analysis, only the three best performing textual sentiment measures were used. For each of the turning points, the forecast has been performed individually. All series have been estimated until three months before the occurrence of the turning point and then the next six periods have been predicted.

Figure 5:52 illustrates the results of the forecast for the three different models at the time of the three different turning points for the all properties *MSCI* series. For the first turning point, the *BING* model reacts one month before the positive growth sets in. Also, the *AFINN* model decreases during the positive growth period. For the second turning point, again only the *BING* model shows a constant increase in the course of the negative growth period. The last turning point does not provide sufficient trends of the three series.

Figure 5:52 - Turning point predictions, MSCI all properties (100,000)



Note 5.116: The three graphs above illustrate the development of the forecast of the textual sentiment indicators during the occurrence of the turning points. The dependent variable in this analysis is the MSCI all properties all assets series.

For the all office series the results have been slightly improved. Table 5:65 illustrates the forecast evaluation for the three models at the three turning points. The reader should keep in mind that the third turning point occurred a couple of months prior to the all properties series. Both the *AFINN* and the *BING* model have a negative sign for the first turning point, while the remaining models stay positive.

The mean squared error results are surprising, given the results of the Diebold Mariano test. The *BING* model is unable to outperform any of the other two models for the second and third turning points. The lowest mean squared error is reached by the *AFINN* model at the first turning point with 0.140.

Similar to before, the results of Theil's U1 increase after the first turning point. Here again, the *AFINN* model has the smallest value of 0.288 in comparison. Comparing with the naïve forecast, all models still produce better results.

Table 5:65 - Forecast evaluation for the three turning points - MSCI all offices (100,000)

	First turning point 2009m8			Second turning point 2012m1			Third turning point 2013m4		
Measures of forecast accuracy	AFINN	BING	*MAXENT* (equal articles)	AFINN	BING	*MAXENT* (equal articles)	AFINN	BING	*MAXENT* (equal articles)
Mean forecast error	-0.070	-0.299	0.114	0.137	0.247	0.280	0.215	0.333	0.215
Mean absolute error	0.359	0.362	0.478	0.592	0.604	0.495	0.368	0.411	0.480
Mean squared error	0.140	0.235	0.248	0.378	0.457	0.325	0.194	0.285	0.277
Root mean squared error	0.375	0.484	0.498	0.614	0.676	0.570	0.440	0.534	0.526
Theil's U1	0.288	0.317	0.452	0.563	0.655	0.612	0.424	0.586	0.530
Theil's U2	0.530	0.685	0.705	0.869	0.956	0.806	0.623	0.755	0.744
C-statistic	-0.718	-0.529	-0.503	-0.244	-0.085	-0.349	-0.611	-0.428	-0.445

Note 5.117: The table evaluates the forecast results for the three turning points of the MSCI all properties all offices series. In this analysis, only the three best performing textual sentiment measures were used. For each of the turning points, the forecast has been performed individually. All series have been estimated until three months before the occurrence of the turning point and then the next six periods have been predicted.

Looking at the graphs of the three models in Figure 5:53, it can be seen that only the first turning point with the *AFINN* and *BING* models reveals the expected behaviour of the indicators. During the second and third turning points, the three models remain relatively stable and do not react to the changes in the market.

Figure 5:53 - Turning point predictions, MSCI all offices (100,000)

Note 5.118: The three graphs above illustrate the development of the forecast of the textual sentiment indicators during the occurrence of the turning points. The dependent variable in this analysis is the MSCI all properties all offices series.

SUMMARY

Once more it has become apparent that the reduction of articles in the seed set for the construction of the sentiment indicators lowers the capability of them to predict the market movement. At the same time, however, the focus on the specific use type (e.g. office) has produced better results. This leads to the conclusion that the underlying nature of the articles has been translated into the indicators.

5.6.1.4.5 SUB-CORPUS V: FINANCIAL TIMES

The *Financial Times* is characterized by a high readership of market professionals. Different from other newspapers the magazine's articles are much more directed towards the broader economy. Therefore, they should carry a much more directed market sentiment in comparison. However, given the three previous analyses, I suspected the results would be weak or even insufficient, due to the low number of articles considered in this sub-corpus (11,948 articles).

Table 5:66 illustrates the summary of statistics for the variables used in this trial. On first glance, there are no distinct differences compared to other sub-corpora. Only the extremes are slightly smaller, which is caused by a smaller number of articles.

Table 5:66 - Summary of statistics (FT)

Variable	Obs	Mean	Std. Dev.	Min	Max
All assets all properties (*MSCI*_change of growth rate)	158	0.297	0.459	0.000	1.000
All assets all offices (*MSCI*_change of growth rate)	158	0.272	0.446	0.000	1.000
AFINN	144	0.000	1.000	-3.688	2.926
BING	144	0.000	1.000	-3.000	2.956
NRC	144	0.000	1.000	-5.315	2.659
TM	144	0.000	1.000	-4.694	2.597
SVM (equal articles)	144	0.000	1.000	-3.982	2.395
MAXENT (equal articles)	144	0.000	1.000	-6.270	2.219
RANDOM FOREST (equal articles)	144	0.000	1.000	-6.092	2.683
MAXENT (all articles)	144	0.000	1.000	-3.891	2.652

Note 5.119: The table illustrates the summary of statistics for the probit analysis for the Financial Times sub-corpus.

Table 5:67 shows the results of the Augmented Dickey-Fuller test. Once again none of the eight indicators or the dependent variables shows any sign of unit roots.

Table 5:67 - Augmented Dickey-Fuller Test (FT)

Variable	Test statistics	1% critical value	5% critical value	10% critical value	Obs.
All assets all properties (*MSCI*_change of growth rate)	-3.568	-3.491	-2.886	-2.576	157
All assets all offices (*MSCI*_change of growth rate)	-4.046	-3.491	-2.886	-2.576	157
AFINN	-6.043	-3.496	-2.887	-2.577	142
BING	-5.414	-3.496	-2.887	-2.577	142
NRC	-5.285	-3.496	-2.887	-2.577	142
TM	-4.487	-3.496	-2.887	-2.577	141
SVM (equal articles)	-7.466	-3.496	-2.887	-2.577	143
MAXENT (equal articles)	-10.032	-3.496	-2.887	-2.577	142
RANDOM FOREST (equal articles)	-6.554	-3.496	-2.887	-2.577	141
MAXENT (all articles)	-6.775	-3.496	-2.887	-2.577	142

Note 5.120: The table illustrates the results of the Augmented Dickey-Fuller Test. All test-statistics are above the critical values at a 1% level.

5.6.1.4.5.1 PROBIT MODEL RESULTS (FT)

In Table 5:68 the probit regression results for the all properties *MSCI* converted growth rate are presented. Most of the models enter the regression with one lag. Only the *SVM* has no lag, while the *TM* and the Random Forrest models have two lags.

As expected, the significance of the various indicators has dropped once more. Besides the *AFINN* and the *BING* model, which are both highly significant at the 1% level, only the other two lexicon-based indicators (*NRC* and *TM*) are significant at the 5% level. The remaining sentiment indicators are insignificant. Both the *SVM* and the Random Forrest models show a positive sign for their coefficient, which is unexpected. The corresponding constant coefficients are all highly significant at the 1% level.

Looking at the pseudo-R-squared value, it can be seen that nearly all models reach values below 5%. The only exceptions are the *AFINN* (0.079) and the *BING* (0.119) models, which provide at least a weak explanation to the dependent variable.

These low values go hand in hand with the misclassification of the observation into either one of the categories. The *BING* model reaches a value of 77.78 and the *AFINN* a score of 76.61. All remaining models reach only values slightly above 0.70, which is a sign of a weak classification. All models pass the *Hosmer-Lemeshow* χ^2 test.

Nevertheless, the results for the area under the *ROC* curve show that all supervised learning algorithms produce only slightly better results than 0.50. For the lexicon-based models, the area scores range between 0.627 (*TM*) and 0.726 (*BING*).

To conclude, the results are by far the weakest in this part of the study. This can only be due to the low number of articles in the seed set for the construction of the indicators. However, the fact that the lexicon approach methods remain superior compared to the machine learning algorithms is striking and confirms my previous observations.

Table 5:68 - Probit results: MSCI - all assets all properties (FT)

Dependent Variable *MSCI* all assets all properties		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	Description	*AFINN*_Articles	*BING*_Articles	*NRC*_Articles	*TM*_Articles	Support Vector Machine	Maximum Entropy (1)	RANDOM FOREST	Maximum Entropy (2)
z_*AFINN*_article	Standardized values for the lexicon approach with the *AFINN* lexicon	-0.423***							
		[0.118]							
z_*BING*_article = L,	Standardized values for the lexicon approach with the *BING* lexicon		-0.541***						
			[0.127]						
z_*NRC*_article = L,	Standardized values for the lexicon approach with the *NRC* lexicon			-0.282**					
				[0.113]					
z_tm_article	Standardized values for the lexicon approach with the *TM* lexicon				-0.238**				
					[0.113]				
z_ceqart_*SVM* = L,	Standardized values for the *SVM* algorithm based on the equalized training corpus with 3 categories					0.183			
						[0.116]			
z_ceqart_max	Standardized values for the *MAXENT* algorithm based on the equalized training corpus with 3 categories						-0.053		
							[0.110]		
z_ceqart_rf	Standardized values for the RF algorithm based on the equalized training corpus with 3 categories							0.083	
								[0.116]	
z_callart_max	Standardized values for the *MAXENT* algorithm based on the full training corpus with 3 categories								-0.058
									[0.113]
Constant		-0.593***	-0.612***	-0.573***	-0.566***	-0.558***	-0.549***	-0.551***	-0.550***
		[0.116]	[0.118]	[0.113]	[0.112]	[0.111]	[0.110]	[0.111]	[0.110]
Observations		144	144	144	144	144	144	144	144
Log-likelihood		-80.100	-76.620	-83.770	-84.700	-85.630	-86.810	-86.660	-86.790
LR Chi2		13.650	20.610	6.309	4.450	2.590	0.231	0.521	0.262
Number of lags		1	1	1	2	0	1	2	1
pseudo-R-squared		0.079	0.119	0.036	0.026	0.015	0.001	0.003	0.002
AIC		164.201	157.235	171.538	173.397	175.258	177.616	177.326	177.585
BIC		170.140	163.174	177.478	179.337	181.198	183.556	183.266	183.525
Correctly classified (%)		73.610	77.780	70.140	70.140	70.830	70.830	70.830	70.830
Sensitivity		16.670	95.100	0.000	0.000	0.000	0.000	0.000	0.000
Specificity		97.060	35.710	99.020	99.020	100.000	100.000	100.000	100.000
Hosmer-Lemeshow χ^2		9.790	7.320	8.780	2.490	4.270	8.030	6.940	7.230
Prob > χ^2		0.28	0.502	0.361	0.962	0.831	0.430	0.543	0.512
area under Receiver Operating Characteristic (*ROC*) curve		0.706	0.726	0.652	0.627	0.562	0.556	0.536	0.516

Standard errors in brackets; *** p<0.01, ** p<0.05, * p<0.1

Note 5.121: The table illustrates the probit results for the MSCI, all assets, all properties series. It can be seen that all lexicon-based sentiment indicators remain significant. The AFINN and the BING models are highly significant at a 1% level, while the other two indicators are significant q at a 5% level. The supervised learning indicators are all insignificant. Again, the BING measure does outperform all remaining measures according to the pseudo-R-squared value. As a test data set the Financial Times sub-corpus was used.

Table 5:69 illustrates the results for the all office *MSCI* converted growth rate. All models enter the regression with a lag, while all lexicon-based models and the *MAXENT* I model have two lags.

Unfortunately, this does not improve the significance of additional indicators. Four of the eight indicators remain insignificant (supervised learning indicators). The significance of the four lexicon-based models remains unchanged. For the four significant models (*AFINN*, *BING*, *NRC* and *TM*) the coefficient sign remains negative. All constant coefficients remain highly significant at the 1% level.

The values of the pseudo-R-squared have slightly improved. The highest value is again produced by the *BING* model with 0.139. The results for the classification remain weak. Nearly all models prefer the majority category and fail to distribute the observations accordingly. All models pass the *Hosmer-Lemeshow* χ^2 test.

Regarding the area under the *ROC* curve, the results for the *AFINN* (0.727) and the *BING* (0.749) model have been improved, while the remaining models remain unchanged at a level below 0.70.

This confirms that the focus within the articles on the commercial real estate side provides a better indication of the market when the dependent variable is also directed towards a more specific market.

Table 5:69 - Probit results: MSCI - all assets - all office properties (FT)

Dependent Variable cg_aa_o		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	Description	*AFINN*_Articles	*BING*_Articles	*NRC*_Articles	*TM*_Articles	Support Vector Machine	Maximum Entropy (1)	RANDOM FOREST	Maximum Entropy (2)
z_*AFINN*_article	Standardized values for the lexicon approach with the *AFINN* lexicon	-0.470***							
		[0.122]							
z_*BING*_article = L,	Standardized values for the lexicon approach with the *BING* lexicon		-0.590***						
			[0.133]						
z_*NRC*_article = L,	Standardized values for the lexicon approach with the *NRC* lexicon			-0.298**					
				[0.116]					
z_tm_article	Standardized values for the lexicon approach with the *TM* lexicon				-0.246**				
					[0.116]				
z_ceqart_*SVM* = L,	Standardized values for the *SVM* algorithm based on the equalized training corpus with 3 categories					0.064			
						[0.116]			
z_ceqart_max	Standardized values for the *MAXENT* algorithm based on the equalized training corpus with 3 categories						-0.092		
							[0.111]		
z_ceqart_rf = L,	Standardized values for the RF algorithm based on the equalized training corpus with 3 categories							0.036	
								[0.116]	
z_callart_max = L,	Standardized values for the *MAXENT* algorithm based on the full training corpus with 3 categories								-0.06
									[0.116]
Constant		-0.718***	-0.745***	-0.684***	-0.674***	-0.654***	-0.656***	-0.653***	-0.654***
		[0.121]	[0.124]	[0.117]	[0.115]	[0.113]	[0.113]	[0.113]	[0.113]
Observations		144	144	144	144	144	144	144	144
Log-likelihood		-74.09	-70.62	-78.73	-79.79	-81.9	-81.72	-82.01	-81.92
LR Chi2		15.93	22.88	6.655	4.529	0.305	0.674	0.096	0.267
Number of lags		2	2	2	2	1	2	1	1
pseudo-R-squared		0.097	0.139	0.040	0.027	0.001	0.004	0.000	0.001
AIC		152.179	145.236	161.458	163.584	167.808	167.439	168.017	167.846
BIC		158.119	151.175	167.398	169.524	173.748	173.379	173.956	173.785
Correctly classified (%)		75.690	79.170	73.610	73.610	74.310	74.310	74.310	74.310
Sensitivity		13.510	29.730	0.000	0.000	0.000	0.000	0.000	0.000
Specificity		97.200	96.260	99.070	99.070	100.000	100.000	100.000	100.000
Hosmer-Lemeshow χ^2		12.050	7.000	6.180	3.930	8.710	6.590	7.340	4.510
Prob > χ^2		0.149	0.536	0.627	0.863	0.367	0.582	0.500	0.808
area under Receiver Operating Characteristic (*ROC*) curve		0.727	0.749	0.660	0.637	0.511	0.570	0.503	0.517

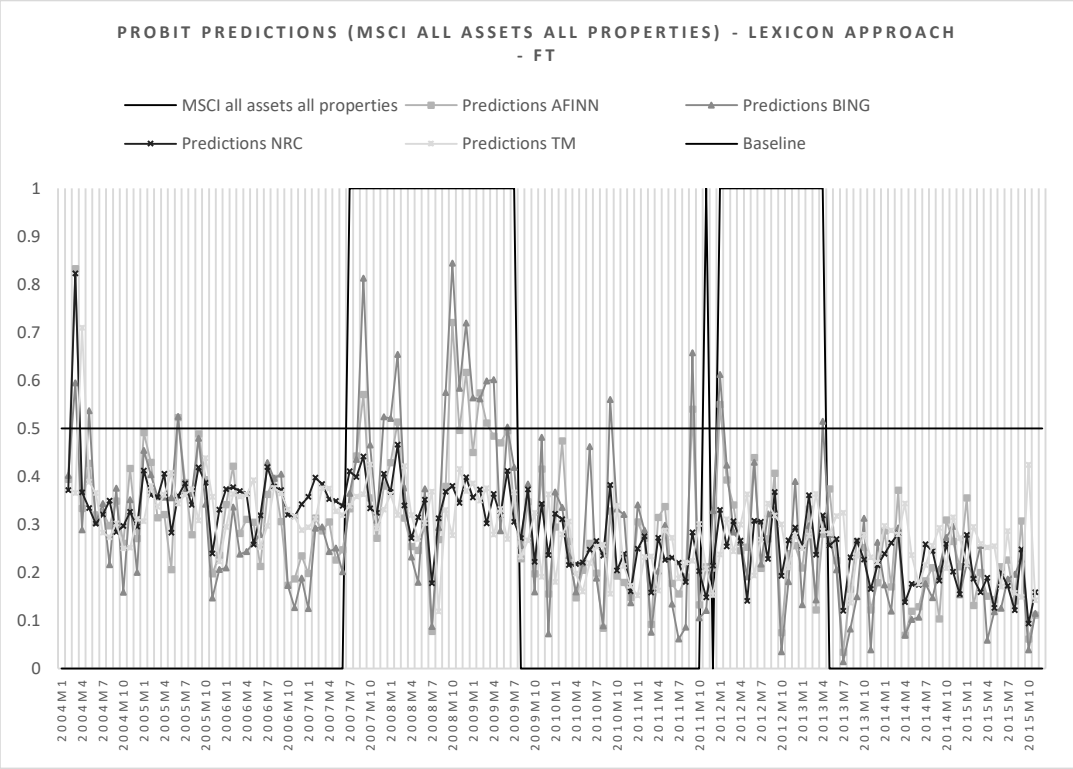
Standard errors in brackets; *** p<0.01, ** p<0.05, * p<0.1

Note 5.122: The table illustrates the probit results for the MSCI, all assets, all offices series. It can be seen that all lexicon-based sentiment indicators remain significant. The AFINN and the BING models are highly significant at a 1% level, while the other two indicators are significant q at a 5% level. The supervised learning indicators are all insignificant. Again, the BING measure does outperform all remaining measures according to the pseudo-R-squared value. As a test data set the Financial Times sub-corpus was used.

5.6.1.4.5.2 PREDICTIONS (FT)

The following two figures illustrate the predictions made by the models for the all properties *MSCI* converted growth rate. Starting with the lexicon approach, Figure 5:54 shows that the two under-performing indicators (*NRC* and *TM*) barely react to the changes in the market. The reaction of the *AFINN* and *BING* model is positive during the first negative growth period (2007m8–2009m7). The correction towards the end is especially picked up by both models. Unfortunately, the models fail to mirror the market path in the subsequent periods.

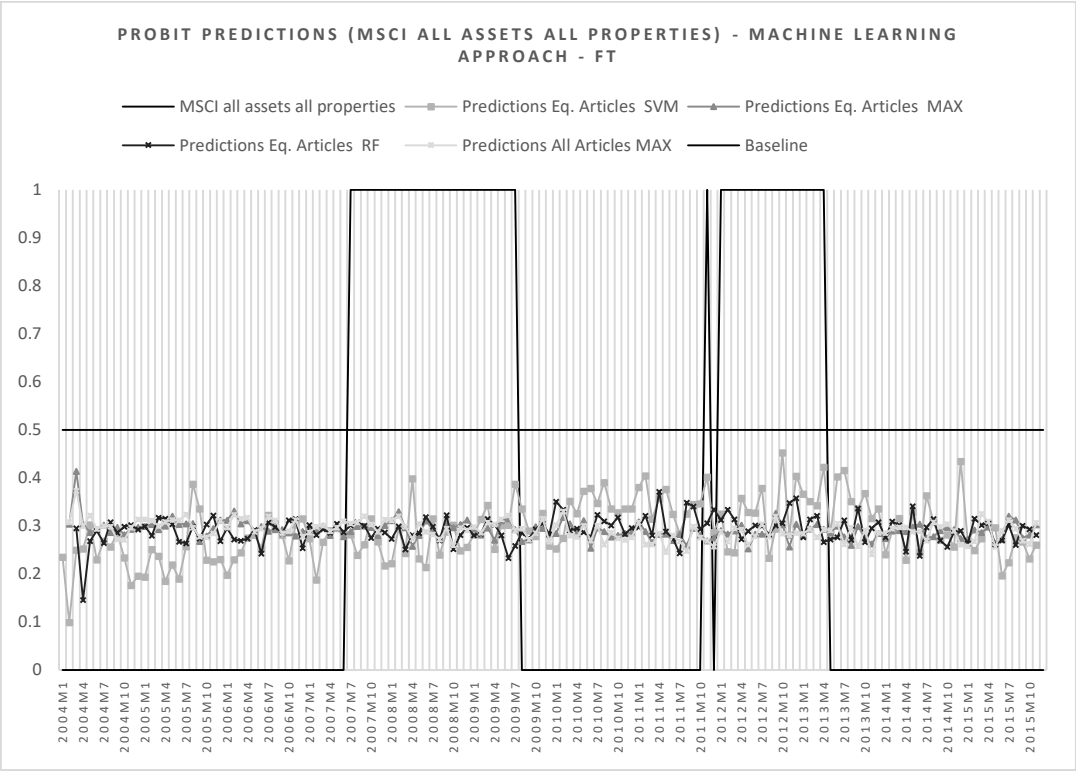
Figure 5:54 - Predictions of the MSCI all properties indicator - lexicon approach (FT)



Note 5.123: The figure illustrates the probit predictions of the four lexicon-based sentiment measures, which have extracted the sentiment from the Financial Times sub-corpus, for the MSCI all assets all properties series.

The results presented in Figure 5:55 only confirm what has been presented in the regression results. The graphs do not resemble the market development, and none of the models picks up any trend.

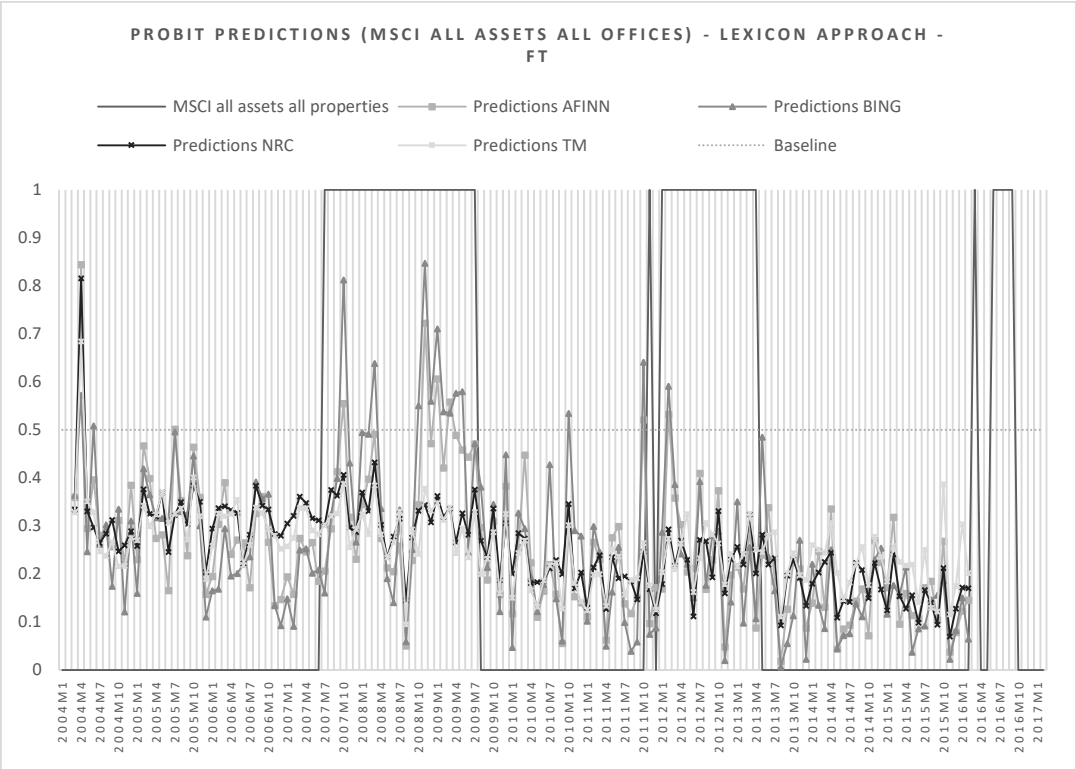
Figure 5:55 - Predictions of the MSCI all properties indicator - machine learning approach (FT)



Note 5.124: The figure illustrates the probit predictions of the four machine learning sentiment measures, which have extracted the sentiment from the Financial Times sub-corpus, for the MSCI all assets all properties series.

Figure 5:56 and Figure 5:57 illustrate the predictions of the eight different indicators for the all office *MSCI* converted growth rate. Both graphs show a slight improvement, at least for the *AFINN* and the *BING* model. Again, these improvements can only be observed for the first period with negative growth.

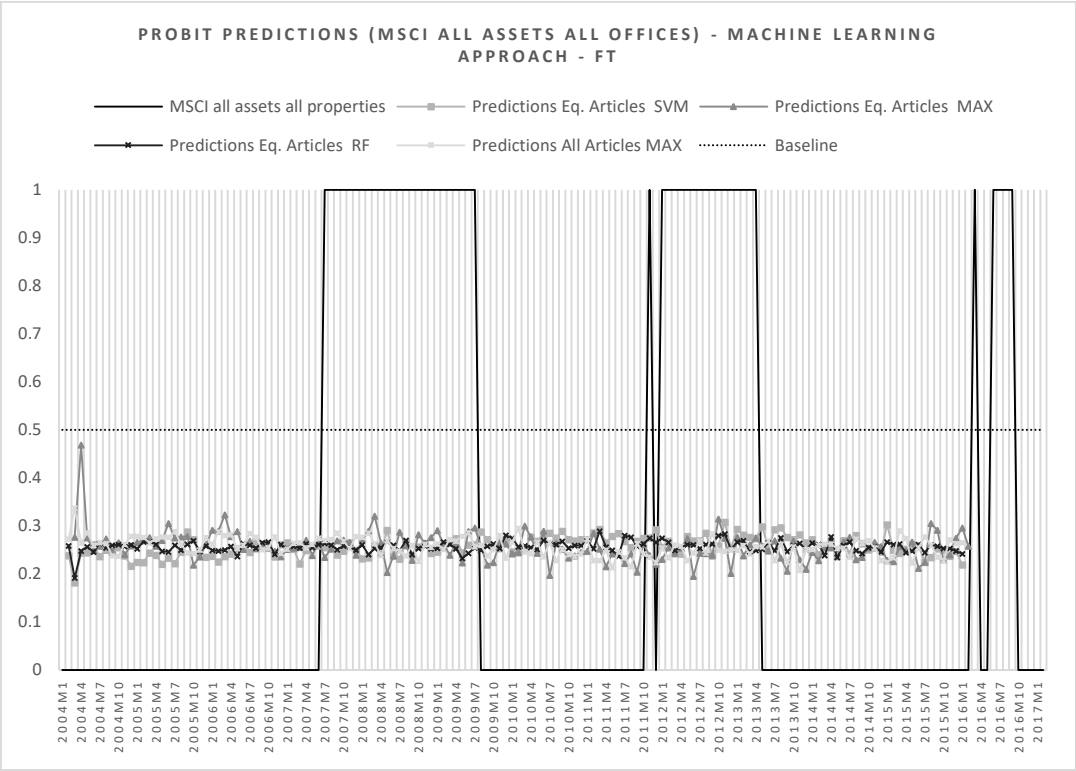
Figure 5:56 - Predictions of the MSCI all offices indicator - lexicon approach (FT)



Note 5.125: The figure illustrates the probit predictions of the four lexicon measures, which have extracted the sentiment from the Financial Times sub-corpus, for the MSCI all assets all offices series.

Figure 5:57 illustrates the predictions of the four FT machine learning indicators. None of the indicators is able to predict any market movement over the course of the testing period.

Figure 5:57 - Predictions of the MSCI all offices indicator - machine learning approach (FT)



Note 5.126: The figure illustrates the probit predictions of the four machine learning measures, which have extracted the sentiment from the Financial Times sub-corpus, for the MSCI all assets all offices series.

5.6.1.4.5.3 DIEBOLD MARIANO TEST (FT)

The Diebold Mariano test results once more confirm what the regression results had suggested. The *BING* model outperforms on an overall level all remaining models for both the all properties (Table 5:70) and the all offices (Table 5:71) series. For the all property series, the MSE of the *BING* model is roughly 0.173 and for the all offices series it is 0.155. Both times, the closest value is again provided by the *AFINN* model.

Table 5:70 - Diebold Mariano Test - MSCI all properties (FT)

	MSE	Difference	S (1)	p-value
BING	0.173			
AFINN	0.185	-0.012	-1.411	0.158
NRC	0.198	-0.025	-1.742	0.082
TM	0.202	-0.028	-1.696	0.090
SVM (equal articles)	0.204	-0.030	-1.091	0.275
MAXENT (equal articles)	0.206	-0.033	-1.536	0.125
RANDOM FOREST (equal articles)	0.207	-0.034	-1.409	0.159
MAXENT (all articles)	0.206	-0.033	-1.537	0.124

Note 5.127: The table illustrates the results of the Diebold-Mariano Test for the MSCI all properties all assets series, for those indicators, which have extracted the sentiment from the Financial Times sub-corpus. The BING series has been used as a reference for the test and all remaining series are evaluated against it.

Table 5:71 - Diebold Mariano Test - MSCI all offices (FT)

	MSE	Difference	S (1)	p-value
BING	0.155			
AFINN	0.166	-0.010	-1.362	0.173
NRC	0.183	-0.027	-1.816	0.069
TM	0.185	-0.029	-1.745	0.081
SVM (equal articles)	0.191	-0.034	-1.459	0.144
MAXENT (equal articles)	0.190	-0.034	-1.640	0.100
RANDOM FOREST (equal articles)	0.191	-0.034	-1.508	0.131
MAXENT (all articles)	0.191	-0.034	-1.594	0.110

Note 5.128: The table illustrates the results of the Diebold Mariano Test for the MSCI all properties all offices series, for those indicators, which have extracted the sentiment from the Financial Times sub-corpus. The BING series has been used as a reference for the test and all remaining series are evaluated against it.

5.6.1.4.5.4 TURNING POINTS (FT)

For the three different turning points, the forecast evaluations are comparable to the 100,000 sub-corpus. Starting again with the all properties series (Table 5:72), it can be seen that all models have a positive mean forecast error, which indicates that they under-predict the dependent variable. Both the *AFINN* (0.177) and the *MAXENT* I model (0.142) produce the lowest values for the first turning point.

As before, the mean squared errors increase over the second turning point and decrease for the last period. The *BING* model (0.222) shows the lowest value for the first turning point. The model further outperforms the other two models consistently for all three periods.

Looking at Theil's U1, it becomes apparent that only the first turning point produces moderate values ranging between 0.451 (*BING*) and 0.503 (*AFINN*). In the subsequent periods, these values increase, especially over the second turning point. Compared to the naïve forecast, all models remain superior.

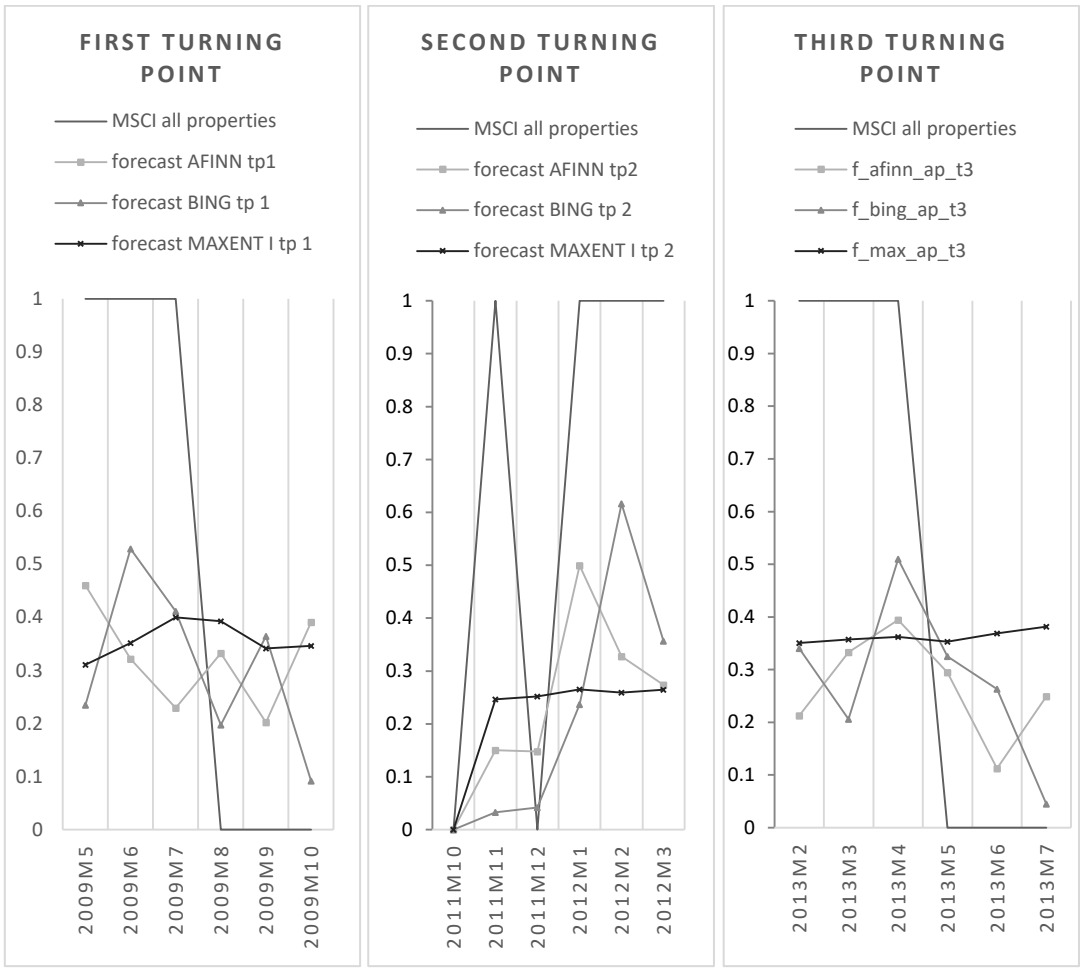
Table 5:72 - Forecast evaluation for the three turning points - MSCI all properties (FT)

	First turning point 2009m8			Second turning point 2012m1			Third turning point 2013m5		
Measures of forecast accuracy	AFINN	BING	*MAXENT* (equal articles)	AFINN	BING	*MAXENT* (equal articles)	AFINN	BING	*MAXENT* (equal articles)
Mean forecast error	0.177	0.195	0.142	0.403	0.421	0.406	0.234	0.218	0.137
Mean absolute error	0.485	0.413	0.503	0.620	0.511	0.576	0.452	0.429	0.505
Mean squared error	0.274	0.222	0.274	0.441	0.366	0.384	0.265	0.247	0.274
Root mean squared error	0.524	0.471	0.523	0.664	0.605	0.619	0.515	0.497	0.524
Theil's U1	0.503	0.451	0.491	0.665	0.677	0.640	0.521	0.486	0.490
Theil's U2	0.741	0.666	0.740	0.813	0.741	0.759	0.728	0.703	0.741
C-statistic	-0.450	-0.555	-0.451	-0.337	-0.450	-0.423	-0.468	-0.505	-0.450

Note 5.129: The table evaluates the forecast results for the three turning points of the MSCI all properties all assets series. In this analysis, only the three best performing textual sentiment measures were used. For each of the turning points, the forecast has been performed individually. All series have been estimated until three months before the occurrence of the turning point and then the next six periods have been predicted.

As before Figure 5:58 presents the graph of the three different models over the course of three different turning points. Given the above-described regression results and the presented forecast evaluations, the graphs are of poor quality. Once more, only the first turning point shows a small resemblance to the market development. In the remaining period, the indicators do not react with enough strength to underlying market development.

Figure 5:58 - Turning point predictions, MSCI all properties (FT)



Note 5.130: The three graphs above illustrate the development of the forecast of the textual sentiment indicators during the occurrence of the turning points. The dependent variable in this analysis is the MSCI all properties all assets series.

Table 5:73 presents the forecast evaluation results of the three models over the three turning point periods for the all offices series. While for the all properties series, the results have been uniform, now the *AFINN* and the *BING* model have a negative sign for the first turning point, meaning that both over predict the dependent variable.

Looking at the mean squared error, the *AFINN* model outperforms the other models for the first and third turning points. The error once again increases over the second period and decreases during the third. For Theil's U1 only the three values in the first period are relatively close to 0, ranging from 0.288 (*AFINN*) to 0.452 (*MAXENT* I).

Comparing the models with a naïve forecast, both Theil's U2 and the C-statistic confirm that all models do better in comparison.

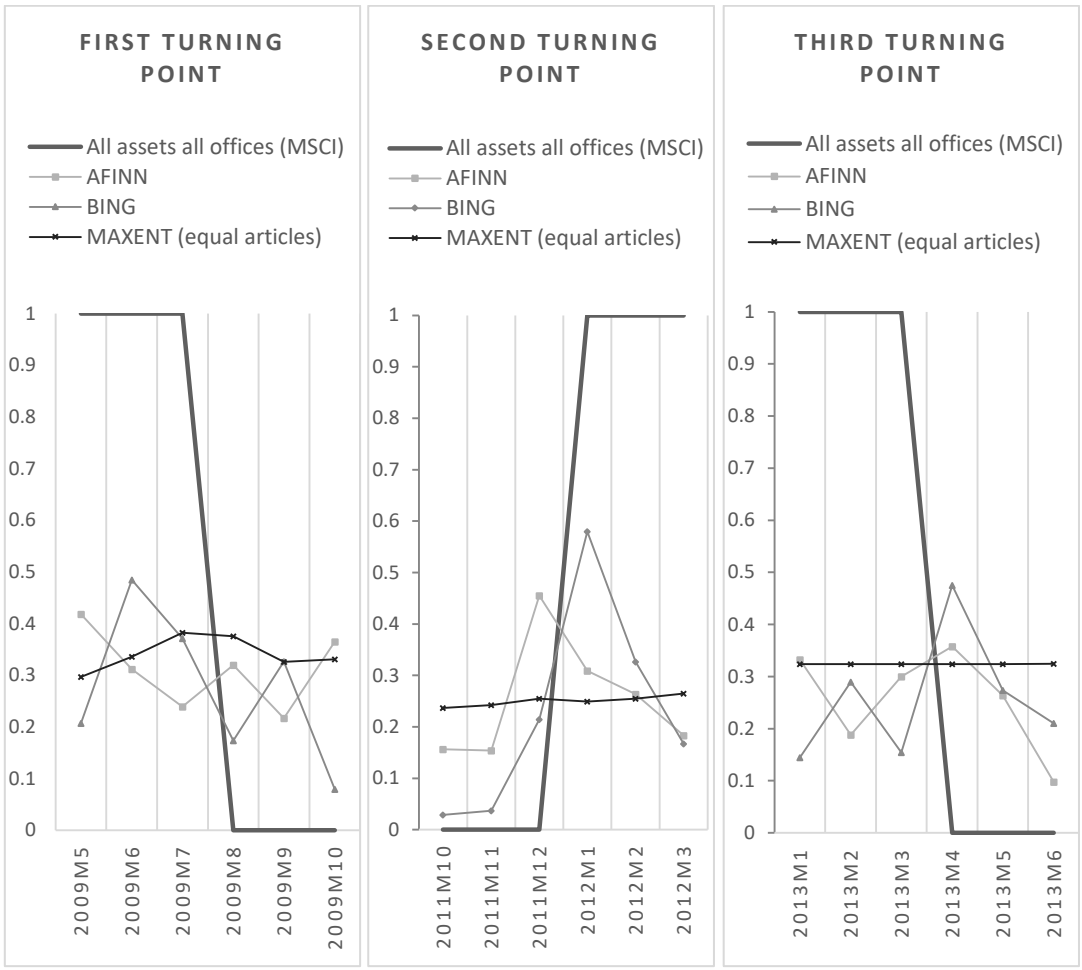
Table 5:73 - Forecast evaluation for the three turning points - MSCI all offices (FT)

.	First turning point 2009m8			Second turning point 2012m1			Third turning point 2013m4		
Measures of forecast accuracy	AFINN	BING	*MAXENT* (equal articles)	AFINN	BING	*MAXENT* (equal articles)	AFINN	BING	*MAXENT* (equal articles)
Mean forecast error	-0.070	-0.299	0.114	0.137	0.247	0.280	0.215	0.333	0.215
Mean absolute error	0.359	0.362	0.478	0.592	0.604	0.495	0.368	0.411	0.480
Mean squared error	0.140	0.235	0.248	0.378	0.457	0.325	0.194	0.285	0.277
Root mean squared error	0.375	0.484	0.498	0.614	0.676	0.570	0.440	0.534	0.526
Theil's U1	0.288	0.317	0.452	0.563	0.655	0.612	0.424	0.586	0.530
Theil's U2	0.530	0.685	0.705	0.869	0.956	0.806	0.623	0.755	0.744
C-statistic	-0.718	-0.529	-0.503	-0.244	-0.085	-0.349	-0.611	-0.428	-0.445

Note 5.131: The table evaluates the forecast results for the three turning points of the MSCI all properties all offices series. In this analysis, only the three best performing textual sentiment measures were used. For each of the turning points, the forecast has been performed individually. All series have been estimated until three months before the occurrence of the turning point and then the next six periods have been predicted.

Looking at the graphs of the models, it can be seen one last time that the first turning point is the only time where the models are able to mirror the market development. The *MAXENT* model, on the other hand, fails to show the required market resemblance (Figure 5:59).

Figure 5:59 - Turning point predictions, MSCI all offices (FT)



Note 5.132: The three graphs above illustrate the development of the forecast of the textual sentiment indicators during the occurrence of the turning points. The dependent variable in this analysis is the MSCI all properties all offices series.

SUMMARY

To conclude, the analysis of the FT indicators has proven my previous observations that the more specific sentiment indicators, which were assumed to perform better, failed to produce sufficient results. While in all cases the performance increased from the general all *MSCI* all properties converted capital growth rate to the all office series, the individual indicators failed to outperform the all articles indicators. This result is somewhat surprising and might have been caused by the number of articles which were included in the sentiment indicator construction.

5.6.1.5 ROBUSTNESS CHECKS

The above-presented results show various things. First, the machine learning algorithms are unable to outperform the more straightforward lexicon-based indicators. Second, within the lexicon-based indicators, both the *AFINN* and especially the *BING* model perform better throughout the whole analysis. And third, the rearrangement of the corpus to a more specific and focused share of the articles, unfortunately, does not lead to an improvement of the indicators. Comparing the five different sub-corpora with each other, the more specific ones produce weaker results compared to the full corpus. The only exception is the 100,000 corpus, which produces weaker results than the complete corpus, but much better results compared to the other three. This confirms my initial assumption, that main newspapers are able to influence the market more due to their more extensive coverage.

The conclusion which I have drawn from this is that the number of articles plays a vital role in the extraction of the sentiment. Given the fact that the overall corpus analysis has produced sufficient results and that the articles have been collected with a focus on the commercial real estate market, the test can be seen as a success, especially if we consider the improvement which has been observed by switching from the more general all properties series to the all offices series.

In the following, I have selected three different robustness tests. The focus is set on different things, but mainly to check whether the indicators can hold their promising results against other types of sentiment indicators and further to test how they react to a different set of dependent variables.

Given the poor performance and to validate my conclusion that the number of articles plays a vital role in sentiment construction, the first test is compiled to see if the indicators perform differently when the underlying dependent variable is more directed to the sentiment indicator. The no housing, the London and the FT indicators should be applied to a more specific dependent variable.

Therefore, in the first test, I use again the three superior models from the above analysis (*AFINN*, *BING* and *MAXENT* I (equal training corpus)) and apply them to another set of two *MSCI* indicators. One concerns the London City Office market, and another concerns the London Mid-Town and West End office market. The idea is to see if the textual sentiment indicators are able to show a stronger and more powerful relationship to the new underlying dependent variable. Both *MSCI* capital growth rates have been again modified into a binary series, with 1 equal to negative growth.

The second robustness check will verify if the newly constructed textual sentiment indicators produce sufficient results in comparison to the direct sentiment measures. I utilize the RICS survey measures, which will enter the same probit model as the textual sentiment indicators.I assume, that the newly constructed measures should perform equally well since they are based on a more straightforward approach.

The last robustness test will put the constructed textual sentiment indicators in the broader picture of this thesis, where I will compare the newly constructed indicators to the previously used indicators. Following my theory, the textual sentiment indicators should outperform the macroeconomic, the office specific and the Google Trends indicators from Chapter 3. This will be tested in a yield model framework.

5.6.1.5.1 ROBUSTNESS CHECK 1: APPLICATION OF THE TEXTUAL SENTIMENT INDICATORS TO MORE LONDON SPECIFIC SERIES

The two new dependent variables of the *MSCI* series provide a more targeted view of the London commercial real estate market. I hope that the effect, which I had observed before, that the results improve by switching from the all properties series to the all office series remains present.

Since not all models have provided sufficient results, I will only compare the results for the *AFINN*, the *BING* and the *MAXENT* I models. These models have previously shown that they are robust to the changing circumstances of the models. The regression results for the different methods are presented in the Appendix (Table 8:31–Table 8:35).

In general, it can be seen that the *BING* model remains superior compared to the other two models. However, all three models show significant improvements in their performance compared to the previous analysis. Primarily, the indicators based on the focused sub-corpora (no housing, London, 100,000 and the FT) show much higher pseudo-R-squared values throughout all three models.

Table 5:74 illustrates the regression results of the three models against all four dependent variables. Panel 1 shows the results for the all properties *MSCI* converted capital growth rate. Panel 2 shows the results for the all office series. These results are a centralization of the previous results. Panels 3 and 4 show the results for the two new dependent variables, the *MSCI* all City of London offices and the *MSCI* capital growth rate for the offices in Mid-Town and West End.

In general, most coefficients are highly significant at the 1% level and carry a negative sign. Only some coefficients for the *MAXENT* I model in the FT sub-corpus are insignificant, and in other instances, the coefficients are only significant at the 5% or 10% level.

Looking at the various pseudo-R-squared values, it can be seen that the leading indicators based on all articles perform reasonably well throughout the four tests, with a slightly better performance towards the second panel (*MSCI* all offices capital growth rate), where the *BING* model peaks in terms of pseudo-R-squared at 0.345. Since the articles have been selected regarding the commercial real estate market, they should have a stronger exposure to the all office category. The *MSCI* all properties capital growth rate incorporates various other factors, such as multiple regions within the U.K. and other use types such as logistics or retail.

Table 5:74 - Comparison of the regression results for the AFINN, BING and MAXENT I models

		(1) *MSCI* all properties capital growth rate			(2) *MSCI* all offices capital growth rate			(3) *MSCI* all city offices capital growth rate			(4) *MSCI* all offices in London Mid-Town & West End capital growth rate		
		AFINN	BING	MAXENT I	AFINN	BING	MAXENT I	AFINN	BING	MAXENT I	AFINN	BING	MAXENT I
All articles	Coefficient	-0.706***	-0.898***	-0.679***	-0.794***	-1.025***	-0.756***	-0.731***	-0.764***	-0.664***	-0.633***	-0.678***	-0.691***
	Standard errors	[0.135]	[0.149]	[0.134]	[0.142]	[0.164]	[0.139]	[0.143]	[0.138]	[0.137]	[0.132]	[0.139]	[0.150]
	Pseudo-R-square	0.195	0.281	0.179	0.243	0.345	0.221	0.218	0.241	0.183	0.196	0.212	0.203
No housing	Coefficient	-0.591***	-0.743***	-0.212*	-0.687***	-0.860***	-0.255**	-0.703***	-0.900***	-0.311**	-0.698***	-1.301***	-0.357***
	Standard errors	[0.129]	[0.150]	[0.113]	[0.139]	[0.166]	[0.117]	[0.149]	[0.169]	[0.123]	[0.149]	[0.248]	[0.133]
	Pseudo-R-square	0.140	0.189	0.021	0.182	0.244	0.030	0.189	0.272	0.045	0.214	0.437	0.061
London	Coefficient	-0.457***	-0.607***	-0.241**	-0.512***	-0.658***	-0.296**	-0.741***	-0.815***	-0.672***	-1.141***	-1.051***	-0.471***
	Standard errors	[0.133]	[0.147]	[0.120]	[0.135]	[0.149]	[0.121]	[0.163]	[0.164]	[0.181]	[0.216]	[0.190]	[0.129]
	Pseudo-R-square	0.089	0.140	0.028	0.114	0.168	0.042	0.203	0.245	0.141	0.391	0.397	0.121
100,000	Coefficient	-0.614***	-0.769***	-0.490***	-0.664***	-0.794***	-0.574***	-0.706***	-1.053***	-0.810***	-0.855***	-1.237***	-0.977***
	Standard errors	[0.127]	[0.143]	[0.122]	[0.130]	[0.144]	[0.129]	[0.134]	[0.173]	[0.148]	[0.159]	[0.205]	[0.176]
	Pseudo-R-square	0.156	0.217	0.104	0.186	0.239	0.139	0.214	0.363	0.246	0.301	0.478	0.34
Financial Times	Coefficient	-0.423***	-0.541***	-0.053	-0.470***	-0.590***	-0.092	-0.576***	-0.697***	-0.204*	-0.607***	-0.827***	-0.171
	Standard errors	[0.118]	[0.127]	[0.110]	[0.122]	[0.133]	[0.111]	[0.136]	[0.151]	[0.118]	[0.144]	[0.173]	[0.120]
	Pseudo-R-square	0.079	0.119	0.001	0.097	0.139	0.004	0.139	0.176	0.021	0.162	0.244	0.016

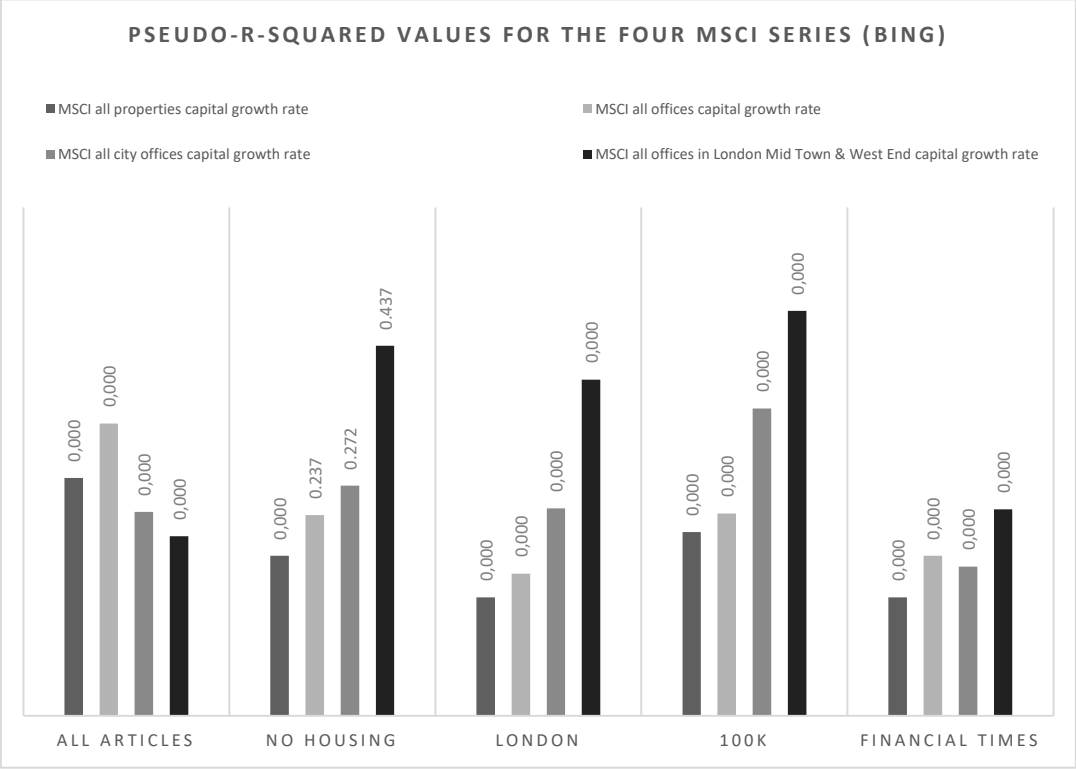
Note 5.133: The table illustrates the coefficient magnitude and significance of the three selected sentiment measures. For each of the 60 probit regressions, the pseudo-R-square value is also presented. Columns 1 and 2 replicate the results from the initial analysis with the MSCI all properties and all offices series. Columns 2 and 3 report the new probit results for the MSCI all city offices and the London Mid-Town and West End series. The bold figures within the pseudo-R-square rows, display the superior models in comparison of the four models.

Looking at the other indicators created with the smaller and more focused sub-corpora, the results for the two panels 1 and 2 are quite weak in comparison. My argument that the poor performance of the indicators is caused by the small number of articles during the construction has not been confirmed.

On the contrary, the indicators outperform the all articles indicators when it comes to a more directed dependent variable. This finding confirms my initial hypothesis that an indicator based on a directed sub-corpus should perform better since the presented sentiment is much more directed.

Take the *BING* indicator for example (Figure 5:60). Its performance decreased in the first two panels from the all articles (0.281) to the *Financial Times* sub-corpus (0.140). However, when the dependent variable is changed to the *MSCI* offices in Mid-Town and West End (Panel 4) the pseudo-R-squared values increase from all articles (0.212) to the 100,000 sub-corpus with 0.478.

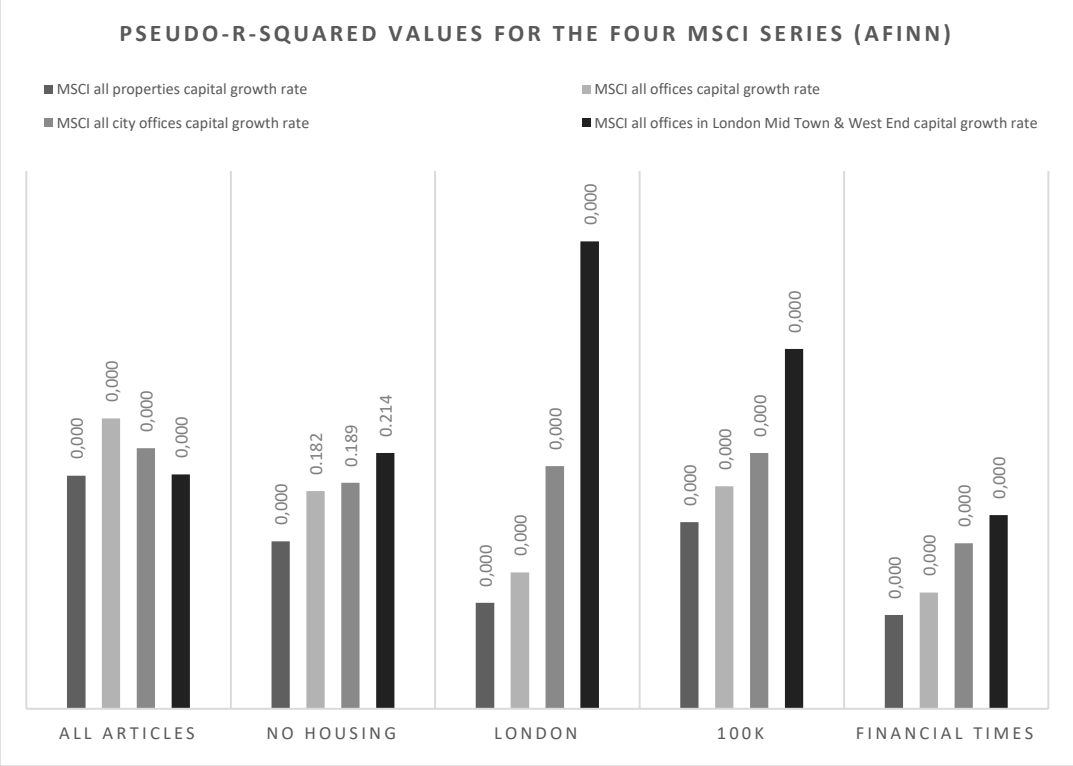
Figure 5:60 - Robustness Check I - BING model – pseudo-R-squared value comparison



Note 5.134: The figure above illustrates the pseudo-R-square values for the BING sentiment induced models for each of the 4 MSCI models and the five different sub-corpora.

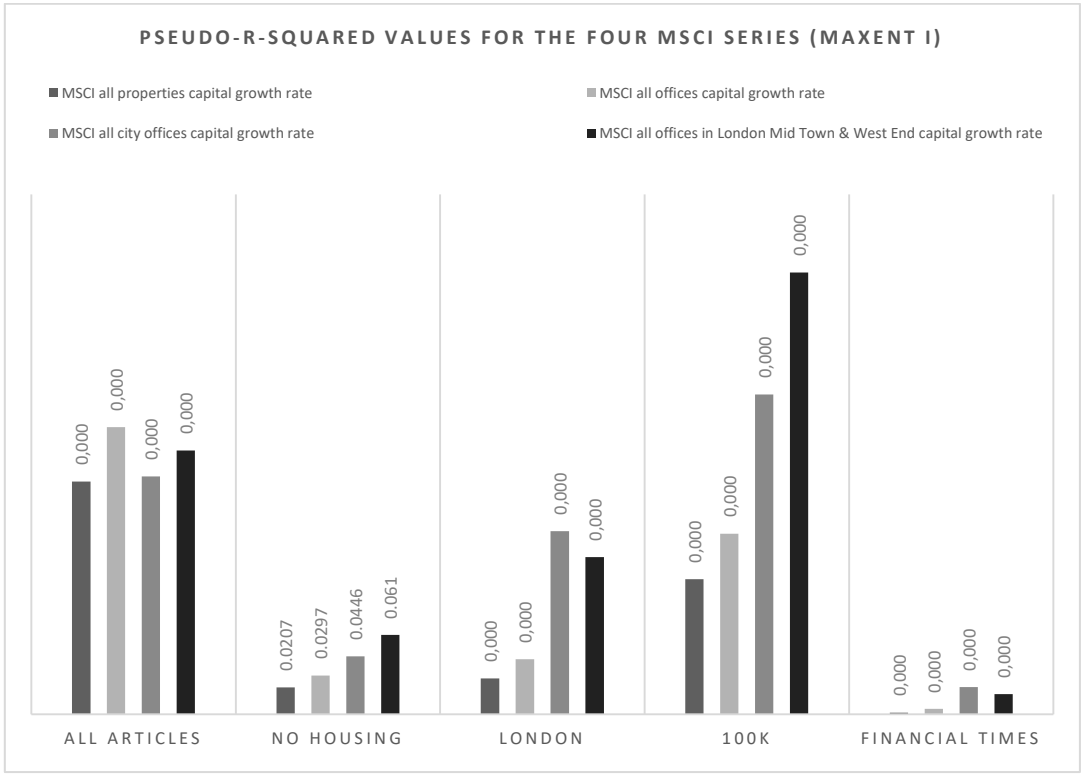
This story holds for the other two approaches as well. For the *AFINN* model (Figure 5:61) the highest pseudo-R-square value is reached by the London specific sub-corpora in the *MSCI* office series for Mid-Town and West End. For the *MAXENT* I model (Figure 5:62), the highest value is also reached by the 100,000 sub-corpora.

Figure 5:61 - Robustness Check I - AFINN model - pseudo-R-squared value comparison



Note 5.135: The figure above illustrates the pseudo-R-square values for the AFINN sentiment induced models for each of the 4 MSCI models and the five different sub-corpora.

Figure 5:62 - Robustness Check I - MAXENT I model - pseudo-R-square value comparison



Note 5.136: The figure above illustrates the pseudo-R-square values for the MAXENT I sentiment induced models for each of the 4 MSCI models and the five different sub-corpora.

Therefore, I am able to conclude that the sentiment is carried in the articles. By rearranging the articles and focusing on more specific market factors, the sentiment becomes clearer. An indicator constructed with all articles might cover the full market; however, it also carries noise. This is why the indicator performs generally well but is not superior when it comes to the more specific market sector.

The no-housing sub-corpus has removed noise factors, by excluding articles which discuss the residential market. This has improved the indicator performance. For the *AFINN* model, the no-housing indicator only outperforms the all articles indicator for the Mid-Town and West End series. And the *MAXENT* I model fails to extract a more suitable sentiment from these articles.

Since the last two series are directly linked to London, it is quite satisfying that the London sub-corpus does so well at least for the *AFINN* model, where it reaches its highest value of 0.391. For the *BING* model, this indicator only ranks third, while the *MAXENT* I model does again not benefit from the more focused sub-corpus.

The 100,000 sub-corpus, on the other hand, provides for the *BING* and the *MAXENT* I model the highest values. That somehow confirms my assumption that a focus on the mainstream newspapers might be enough to extract the market sentiment.

The last idea that market participants read the *Financial Times* and take their information from this newspaper could only be confirmed to a limited extent, given the fact that the indicator ranks fourth for the *BING* and fifth for the *AFINN* model. Nonetheless, the *Financial Times* is already included in the 100,000 sub-corpus.

5.6.1.5.2 ROBUSTNESS CHECK 2: COMPARISON BETWEEN THE RICS SURVEY MEASURES AND THE SUPERVISED LEARNING MEASURES IN A PROBIT MODEL

In a second try, I will apply the two RICS series to the above-used probit model for the *MSCI* office series for the London Mid-Town and West End market. The literature has suggested that sentiment indicators, which are based on survey data, are superior to other sentiment proxies. I expect, therefore, the two RICS series to perform exceptionally well in comparison to the other three models. Since the RICS data is only available on a quarterly basis, we need to use the quarterly measures. Given the above-presented results, I will use the *BING* and *MAXENT* I *100,000 indicators* as well as the *AFINN London indicator* for this comparison.

Table 5:75 illustrates the probit model results for the quarterly analysis. All five sentiment indicators have a negative sign and are highly significant at the 1% level. None of the indicators has entered the model with any lag. As expected, the two RICS sentiment series perform extremely well. Both reach pseudo-R-squared values above 0.40, which is only achieved by the *BING* model. While the *AFINN* (0.365) and the *MAXENT* I (0.262) model are both outperformed by both series, the *BING* (100,000) model (0.457) is able to perform better than the office RICS measure (0.447). However, it fails to outperform the general RICS measure which reaches the highest pseudo-R-squared value with 0.468.

Table 5:75 - Probit model RICS vs best indicators

					MSCI office Mid-Town & West End
	(1) RICS office	(2) RICS general market	(3) *AFINN*_articles (London)	(4) *BING*_Articles (100,000)	(5) Maximum Entropy (1) (100,000)
z_rics_off	-1.279***				
	[0.388]				
z_rics_all		-1.551***			
		[0.531]			
z_*AFINN*_article (London)			-1.318***		
			[0.452]		
z_*BING*_article (100,000)				-1.358***	
				[0.396]	
z_ceqart_max (100,000)					-0.900***
					[0.297]
Constant	-1.080***	-1.136***	-0.860***	-1.217***	-1.074***
	[0.296]	[0.307]	[0.280]	[0.311]	[0.272]
Observations	44	44	37	43	43
Log likelihood	-13.05	-12.55	-13.03	-11.97	-16.28
LR Chi2	21.07	22.06	15	20.18	11.57
Lag	0	0	0	0	0
pseudo R-squared	0.447	0.468	0.365	0.457	0.262
AIC	30.092	29.105	30.054	27.940	36.550
BIC	33.660	32.673	33.276	31.462	40.073
Correctly classified (%)	90.910	90.910	81.080	83.720	81.400
Sensitivity	100.000	100.000	44.440	44.440	33.330
Specificity	60.000	60.000	92.860	94.120	94.120
Hosmer-Lemeshow χ^2	10.440	10.370	10.520	5.490	7.800
Prob > χ^2	0.235	0.240	0.231	0.704	0.453
area under Receiver Operating Characteristic (*ROC*) curve	0.900	0.894	0.877	0.909	0.882

Standard errors in brackets *** p<0.01, ** p<0.05, * p<0.1

Note 5.137: The table above reports the probit results for the five different probit regressions with the two direct sentiment measures and the three constructed textual sentiment measures. For the textual sentiment measures, the AFINN indicator from the London focused corpus has been used. The BING and the MAXENT I indicators are both taken from the 100,000 focused corpus. As a dependent variable, the MSCI office Mid-Town and West End series has been used. All five series have been transformed into a quarterly series.

As expected, the survey-based measures performed reasonably well against the constructed sentiment measures. The fact that the *BING* model has outperformed at least one of them and has only produced slightly worse results than the other, is quite promising.

Reminding the reader of the fact that the survey-based measures are costly in their construction should provide sufficient argument for the textual sentiment indicators being preferred.

5.6.1.5.3 ROBUSTNESS CHECK 3: COMPARISON TO THE MACROECONOMIC SENTIMENT INDICATORS AND TEXTUAL SENTIMENT INDICATORS FROM THE PREVIOUS PARTS

The fourth robustness check is designed to place this chapter within the broader picture of the whole thesis. To justify whether the constructed textual sentiment indicators perform better than the previously used indicators, I will apply them in a basic yield regression model for the London market.

The textual sentiment indicators will compete against the macroeconomic sentiment, the office specific and the Google Trends measure, as well as against the textual sentiment indicator based on the market reports from the second part of this thesis.

Given the performance of the newly constructed textual sentiment and machine learning based indicators, I assume that they should perform at least as well as the office specific indicator, which has been superior in previous tries. However, in comparison to the remaining indicators, the lexicon-based approaches should be able to outperform them.

For the *BING* and the *MAXENT* I model I will use the *100,000 indicator,* and for the *AFINN* model, I will use the *London specific indicator*.

For this test, I will recycle the standard yield model from section 3.6.2.

Table 5:76 - Robustness check 3 - sentiment indicators within a standard yield model

Dependent variable: Office yield for London West End	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
LABELS	Standard yield model (Hendershott)	Standard yield model with Macroeconomic Sentiment	Standard yield model with Office sentiment	Standard yield model with Google Trends	Standard yield model with Textual sentiment (market reports)	Standard yield model with *AFINN* (London)	Standard yield model with *BING* (100,000)	Standard yield model with *MAXENT* I (100,000)
Office rent four quarter moving average	2.800**	3.680**	-0.395	10.582***	13.460***	1.037	-0.372	1.893
	[1.341]	[1.511]	[0.909]	[1.321]	[2.949]	[1.107]	[0.874]	[1.131]
Ten years - Government bond rate	0.376***	0.312***	-0.039	0.062	0.560***	0.099	0.137**	0.192***
	[0.065]	[0.069]	[0.067]	[0.063]	[0.122]	[0.065]	[0.051]	[0.065]
Risk premium	0.078***	0.01	-0.052**	0.097***	0.117***	0.027	0.027*	0.054***
	[0.019]	[0.024]	[0.021]	[0.016]	[0.035]	[0.021]	[0.014]	[0.018]
Macroeconomic Sentiment		-1.276***						
		[0.410]						
Office Sentiment			-0.779***					
			[0.081]					
Google Trends				-1.128***				
				[0.138]				
London CRE Market Reports					-0.747			
					[0.676]			
AFINN (London)						-0.548***		
						[0.070]		
BING (100,000)							-0.562***	
							[0.061]	
MAXENT I (100,000)								-0.425***
								[0.082]
Constant	3.000***	2.401***	5.402***	4.282***	2.665***	4.232***	3.984***	3.708***
	[0.266]	[0.342]	[0.355]	[0.268]	[0.719]	[0.304]	[0.205]	[0.268]
Observations	44	43	42	43	24	37	43	43
R-squared	0.276	0.484	0.754	0.697	0.624	0.708	0.782	0.604
adjusted R-squared	0.222	0.43	0.728	0.665	0.545	0.672	0.759	0.563
Number of Lags		0	1	1	4	0	0	0
AIC		58.289	25.329	35.395	31.507	28.923	21.207	46.877
BIC		67.095	34.018	44.201	37.398	36.977	30.013	55.683
F-statistic	11.830	10.920	34.910	38.790	11.720	21.390	33.270	13.020
Prob > F	0	0	0	0	0.000	0	0	0
degrees of freedom	40	38	37	38	19	33	38	39

Robust standard errors in brackets *** p<0.01, ** p<0.05, * p<0.1

Note 5.138: The table above represents the main comparison of all sentiment indicators from this thesis. For comparison reasons, the textual sentiment indicators have been transformed into a quarterly series. The comparison is performed on the standard yield model from chapter 3. Columns one to four represent the indirect sentiment indicators from chapter 3. Column five applies the textual sentiment

indicator from chapter four, London CRE Market Reports. The remaining columns use the three newly constructed textual sentiment indicators from this chapter. All sentiment induced models have outperformed the base model.

Table 5:76 illustrates the results for the eight different models. Model 1 is the base model, with no sentiment measure. It can be seen that all variables are at least statistically significant at the 5% level. The base model reaches an R-squared value of 0.276.

Looking at the seven sentiment specific models, it becomes apparent that all models have the expected negative coefficient, which is highly significant at the 1% level, except for the textual sentiment indicator based on market reports, which has failed to produce a significant coefficient. Some indicators enter the model lagged. The number of lags has been estimated with the help of the AIC.

Comparing the R-squared values, it can be seen that all sentiment induced models outperform the base model. Even more satisfying is the fact that the *BING* (100,000) model reaches the highest adjusted R-squared value with 0.759, followed by the office specific measure (0.728).

SUMMARY

To conclude, the regression results have proven the superiority of the newly constructed sentiment measures. Applying the *BING* (100,000) measure to the standard yield model has shown that the lexicon approach is suitable for various applications. Compared to the second-ranked office specific measure, the *BING* model is more straightforward and only relies on textual data, while the office measure needs real estate specific information, which is published *ex-post* to the market development.

5.6.2 DEVELOPMENT OF A DIFFERENT TRAINING DATASET USING THE LEXICON APPROACH

A central issue of the above-displayed results is the unknown quality of the final textual sentiment values. No knowledge about the news corpus (test corpus) prior to the analysis is present. The generated labels of the above-displayed analysis have to be accepted as they are. Since no comparison regarding the quality can be made, the output has left room for doubt.

Therefore, this chapter will combine the two previously used methods. The lexicon approach is a straightforward method for labelling a corpus. Using the wordlists, even a large corpus of articles can be classified in a relatively short time. Wordlists have further been proven

in multiple studies as a useful method. In this section, I am going to use these advantages of the lexicon approach to annotate a newly constructed training corpus. The training corpus is then used to train a set of new classifiers which will be used to label the test corpus. The test is performed only on a minor share of the initially collected testing corpus.

Out of the following reasons the FT sub-corpus is used for this analysis. In comparison to the other sub-corpora, the number of FT articles related to real estate remained stable over the whole testing period (see Figure 5:1). Further, this low number of articles in the testing corpus reduces the computation time dramatically when the newly trained algorithms are applied to it.

This approach has been used before by other researchers such as Fang et al. (2011) and Mudinas et al. (2012), who have labelled their corpus with the help of sentiment lexica or used the lexica themselves to train their algorithms with them, such as He and Zhou (2011). The advantage of this approach is that a fast and straightforward analysis of the corpus is possible. Further, the possibility of comparing the given labels of the lexicon approach with the generated labels from the supervised learning algorithms can be seen as a significant improvement upon the previously used labelling process with the *Amazon* book reviews.

Another motivation for this approach is the published work of Augustyniak et al. (2014). The authors state that the use of the lexicon approach is still favourable since supervised learning approaches barely outperform these easy and flexible methods, which only rely on wordlists. So, the question arises, what additional value can be provided by supervised learning methods? If their performance is similar to the basic lexicon approach, then it is unclear why scholars should proceed with supervised learning algorithms for sentiment extraction, given the fact that their development is somewhat time-consuming and complicated in the calculation.

Using either the book reviews or the wordlists as the underlying source for the training of the algorithms leads ultimately to the adoption of a biased structure or pattern. If a classifier is trained with a text, which has been annotated initially with the help of a lexicon, then the algorithm incorporates the characteristics of the different lexica. However, these biases are probably much stronger in the case of book reviews. It is fair to say that the algorithms try to reproduce a pattern in the testing (unknown) corpus, which is similar in nature to the training corpus, which mainly relies on the lexicons.

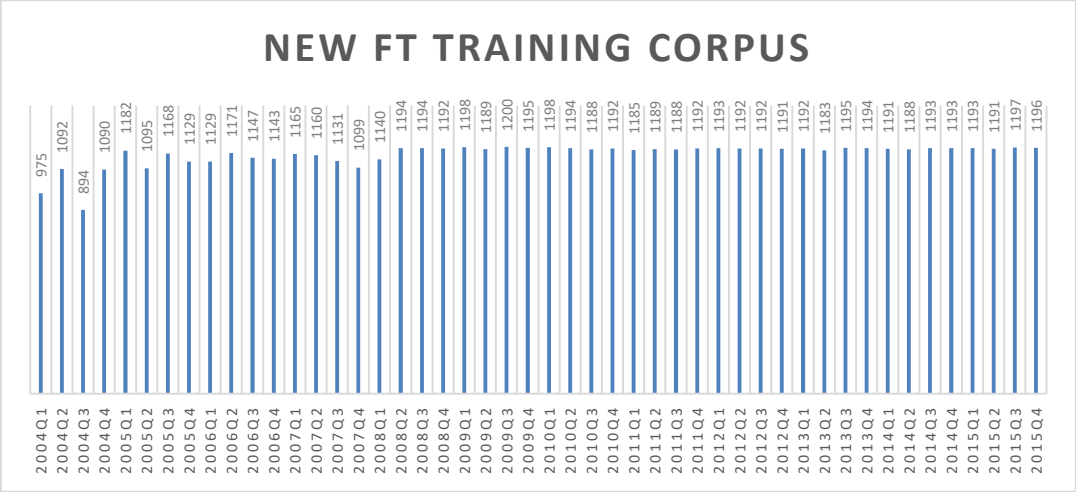
The final classification, however, is performed on the full text of the articles and not on the individual words of the lists. Therefore, it might be possible that the algorithms search for a

more profound pattern, which remains hidden to the human mind. Consequently, there might be a chance that the classifiers not only mirror the lexicons but also find hidden structures in the test corpus which will influence the final classification of a specific document.

I have collected a second dataset of FT news articles for the same period, 2004q1–2015q4. Different to the initial approach (*Amazon* reviews), the newly collected articles are similar in structure and wording to the test dataset. I also assume that the classifiers trained on a similar dataset should be more suitable for extracting the inherent sentiment in the test dataset.

Besides the restriction to use only U.K. related FT articles, I have not filtered for any other options during the collection process on ProQuest News & Newspapers. On average I have collected more than 350 articles on a monthly basis. The new corpus consists of 55,872 entities and is distributed as shown in Figure 5:63.

Figure 5:63 - New FT training corpus

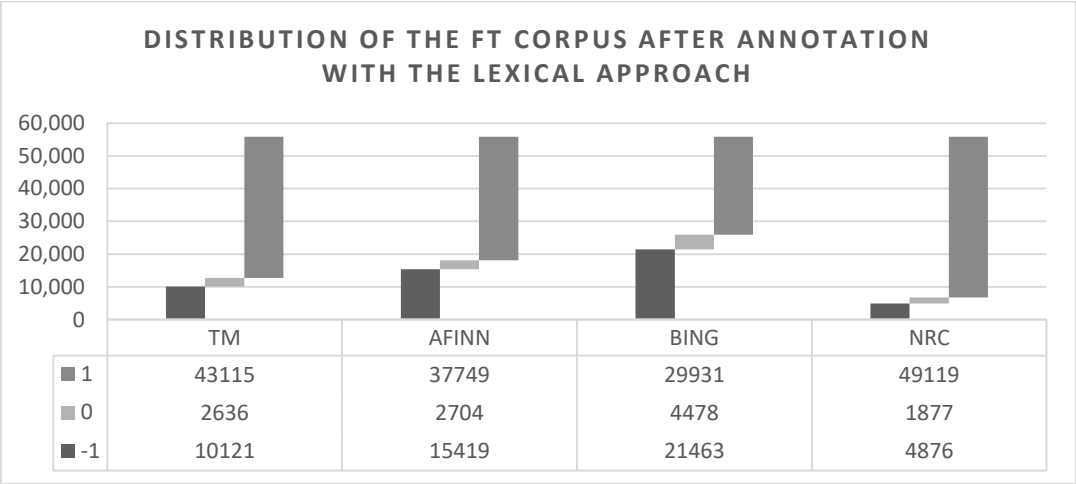


Note 5.139:The figure above illustrates the distribution of the newly collected training corpus, on a quarterly level.

The figure illustrates the distribution of the new training corpus over the full testing period. On average, the number of articles over the quarters remains stable; however, from 2004 to 2007 this number differs slightly. The difference for the first six quarters and in 2007q4 seems to come from the total number of published articles. Another reason, as described earlier, is the mismatch between the displayed and provided a number of articles on ProQuest News & Newspapers.

In a second step, the corpus is labelled through the four lexical approaches (*AFINN, BING, NRC* and *TM*). I have transformed each sentiment value into a specific class positive–neutral–negative with its corresponding numerical value (1, 0, –1).

Figure 5:64 - Distribution of the FT corpus over the three different classes



	TM	AFINN	BING	NRC
1	43115	37749	29931	49119
0	2636	2704	4478	1877
-1	10121	15419	21463	4876

Note 5.140: The figure above illustrates the distribution of the FT training corpus after it has been annotated by the four different lexicons. In total 55,872 entities have been labelled by each approach.

After the corpus has been annotated by the four different lexical approaches, it can be seen in Figure 5:64 that all methods put a stronger emphasis on the positive category. The number of articles which have been labelled as neutral is rather small. Only the *BING* method seems to be able to distribute the positive and negative classes more equally. I first thought that a reason for this classification bias could be found in the structure of the underlying lexicons. Referring back to Table 4:5, it can be seen that the number of positive words is smaller in all four cases, so the only reason for the bias toward the positive category might be found in the words used within the articles. Maybe the articles incorporate more positive words than negative ones.

After the new training corpus was annotated with the lexicons, the corpus was then used to train the different classifiers. Finally, the new classifiers are used to classify the test dataset. In total 11,948 FT articles have been previously used. Training and the test dataset are more or less split into the recommended 80% and 20% share, while in this case, the training dataset is slightly larger at 83.67%.

Finally, the newly constructed sentiment measures will enter another probit model. Due to the good performance of the more focused dependent variables of the MSCI series, I have decided to use the *MSCI* capital growth rate for offices in London Mid-Town and West End.

5.6.2.1 PERFORMANCE ANALYSIS

In this section, the performance of the different algorithms is displayed and compared. Table 5:77 shows the precision, the recall and the corresponding F-score for the different classifiers.

The first significant difference to the above-displayed results is that, due to the lower number of articles in the training dataset, all classifiers could have been calculated. As shown in Table 5:77. However, some algorithms performed better than others. The grey shaded results show where the classifiers were unable to produce sufficient results.

It can be seen that, for the *NRC* trained classifier, five of the nine algorithms failed to produce sufficient results. In three of these cases (*GLMENT*, BAGGING and *NNET)* the reason can be found in the fact that not a single article was sorted into the neutral (0) category. One could assume that the reason for this again can be found in the structure of the underlying lexicon. Yet, the number of neutral words in the *NRC* lexicon is more significant than the words in the other two categories. Therefore, a sufficient number of articles could have been sorted into this category. In the two remaining cases (*RANDOM FOREST* and *TREE*), no article was sorted into either category. Only the *SVM* classifier produced an average precision score of above 0.50.

For the classifiers trained with the *TM* and *AFINN* lexicon, the results are similar. The same algorithms (*GLMENT*, BAGGING, *NNET* and *TREE*) failed to produce acceptable results. Similar reasons apply as before. For both approaches, the *RANDOM FOREST* algorithm produces a higher precision value than any other algorithm. The *AFINN* value of 0.87 actually outperforms any other algorithm in this attempt. Unfortunately, the corresponding recall value is less than 0.50, which states that less than half of the instances have been correctly labelled.

The last applied lexicon is the *BING* lexicon. Six out of nine algorithms are able to produce sufficient results. Again, the *RANDOM FOREST* algorithm performs best in comparison to the other five. On the recall side, the results are again mixed, yet the *RANDOM FOREST* algorithm is able to label more than 50% of the records correctly. The highest recall value was achieved by

the Maximum Entropy classifiers (*MAXENT*) in all four cases, with values as high as 0.58 for the Topic Modelling (*TM*) approach.

Compared to the results above, this story is coherent. In Table 5:6 both *SVM* and *MAXENT* show the highest recall values, above 50%. It seems that these two classifiers are able to outperform the other seven algorithms for the task at hand consistently.

Table 5:77 - Performance analysis – FT news corpus annotated with the sentiment lexicons

Training Model	Class	SVM			MAXENTROPY			GLMENT			SLDA			BAGGING			BOOSTING			RANDOM FOREST			NNET			TREE		
		Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score
NRC_Articles	-1	0.62	0.35	0.45	0.44	0.45	0.44	0.58	0.01	0.02	0.44	0.24	0.31	0.45	0.02	0.04	0.23	0.04	0.07	-	0.00	-	0.37	0.66	0.47	-	0.00	-
	0	0.00	0.00	-	0.08	0.12	0.10	-	0.00	-	0.00	0.00	-	-	0.00	-	0.00	0.00	-	1.00	0.01	0.02	-	0.00	-	-	0.00	-
	1	0.93	0.99	0.96	0.95	0.94	0.94	0.91	1.00	0.95	0.93	0.98	0.95	0.91	1.00	0.95	0.91	0.99	0.95	0.91	1.00	0.95	0.96	0.93	0.94	0.91	1.00	0.95
overall		0.52	0.45	0.71	0.49	0.50	0.49	0.75	0.34	0.49	0.46	0.41	0.63	0.68	0.34	0.50	0.38	0.34	0.51	0.96	0.34	0.49	0.67	0.53	0.71	0.91	0.33	0.95
TM_Articles	-1	0.78	0.72	0.75	0.73	0.70	0.71	0.85	0.21	0.34	0.71	0.46	0.56	0.70	0.17	0.27	0.47	0.25	0.33	0.81	0.05	0.09	0.62	0.80	0.70	-	0.00	-
	0	0.00	0.00	-	0.11	0.10	0.10	-	0.00	-	0.00	0.00	-	-	0.00	-	0.07	0.03	0.04	0.75	0.01	0.02	-	0.00	-	-	0.00	-
	1	0.92	0.97	0.94	0.93	0.94	0.93	0.83	0.99	0.90	0.87	0.97	0.92	0.83	0.99	0.90	0.83	0.93	0.88	0.81	1.00	0.90	0.94	0.92	0.93	0.80	1.00	0.89
overall		0.57	0.56	0.85	0.59	0.58	0.58	0.84	0.40	0.62	0.53	0.48	0.74	0.77	0.39	0.59	0.46	0.40	0.42	0.79	0.35	0.34	0.78	0.57	0.82	0.80	0.33	0.89
*AFINN*_Articles	-1	0.78	0.72	0.75	0.76	0.72	0.74	0.84	0.49	0.62	0.74	0.58	0.65	0.66	0.40	0.50	0.39	0.72	0.51	0.86	0.29	0.43	0.65	0.81	0.72	0.64	0.05	0.09
	0	0.06	0.00	0.00	0.08	0.07	0.07	-	0.00	-	0.00	0.00	-	-	0.00	-	0.19	0.01	0.02	1.00	0.01	0.02	-	0.00	-	-	0.00	-
	1	0.87	0.93	0.90	0.89	0.91	0.90	0.80	0.97	0.88	0.82	0.93	0.87	0.77	0.93	0.84	0.83	0.59	0.69	0.75	0.98	0.85	0.90	0.86	0.88	0.70	0.99	0.82
overall		0.57	0.55	0.55	0.58	0.57	0.57	0.82	0.49	0.75	0.52	0.50	0.76	0.72	0.44	0.67	0.47	0.44	0.41	0.87	0.43	0.43	0.78	0.56	0.80	0.67	0.35	0.46
BING_Articles	-1	0.76	0.78	0.77	0.76	0.77	0.76	0.78	0.67	0.72	0.73	0.70	0.71	0.67	0.55	0.60	0.42	0.87	0.57	0.71	0.65	0.68	0.72	0.72	0.72	0.47	0.55	0.51
	0	0.10	0.00	0.00	0.14	0.09	0.11	-	0.00	-	0.06	0.00	0.00	0.25	0.00	0.00	0.07	0.00	0.00	0.40	0.01	0.02	-	0.00	-	-	0.00	-
	1	0.81	0.89	0.85	0.84	0.86	0.85	0.76	0.91	0.83	0.77	0.87	0.82	0.70	0.86	0.77	0.77	0.32	0.45	0.75	0.87	0.81	0.79	0.87	0.83	0.65	0.64	0.64
overall		0.56	0.56	0.54	0.58	0.57	0.57	0.77	0.53	0.78	0.52	0.52	0.51	0.54	0.47	0.46	0.42	0.40	0.34	0.62	0.51	0.50	0.76	0.53	0.78	0.56	0.40	0.58

Note 5.141: The table illustrates the three performance measures for the nine different algorithms. The results are based on the FT news corpus, which was annotated through the four different sentiment lexicons NRC, TM, AFINN and BING. The assigned sentiment values were modified to numerical values [(-1) 0 1]. A total of 55,872 news articles were used. Each algorithm has been trained on 80% of these reviews, and the displayed results are generated with the remaining 20% as testing values. For each of the algorithms precision, recall and the F-score were calculated on a class level. The "overall" row illustrates the average over the different classes. Grey shaded algorithms have not produced good results; they failed to distribute the entities over the classes.

Table 5:78 - Overall performance comparison between the Amazon book review and the lexical approach

Method	Training model	SVM			MAXENT			GLMENT			SLDA			BAGGING			BOOSTING			RANDOM FOREST			NNET			TREE		
		Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score
AMAZON BOOK REVIEWS	3c_all				0.56	0.51	0.51	0.69	0.35	0.32	0.61	0.39	0.39	0.65	0.40	0.41	0.54	0.42	0.42	0.70	0.39	0.40	0.60	0.45	0.63	0.75	0.33	0.86
	3c_eq	0.62	0.56	0.53	0.58	0.56	0.55	0.73	0.53	0.49	0.64	0.52	0.48	0.57	0.50	0.47	0.58	0.43	0.37	0.62	0.55	0.52	0.65	0.54	0.72	0.50	0.38	0.45
	5s_all				0.39	0.36	0.36	0.45	0.21	0.23	0.41	0.26	0.26	0.37	0.24	0.24				0.45	0.26	0.26	0.48	0.24	0.53	0.58	0.20	0.73
	5s_eq	0.43	0.41	0.42	0.42	0.40	0.40	0.39	0.37	0.38	0.41	0.39	0.40	0.38	0.35	0.32	0.36	0.29	0.21	0.42	0.42	0.41	0.26	0.20	0.13	0.66	0.03	0.17
	Average	0.62	0.56	0.53	0.51	0.48	0.47	0.62	0.37	0.34	0.55	0.39	0.38	0.53	0.38	0.37	0.56	0.42	0.40	0.59	0.40	0.39	0.57	0.41	0.62	0.61	0.30	0.68
LEXICON APPROACH	NRC	0.52	0.45	0.71	0.49	0.50	0.49	0.75	0.34	0.49	0.46	0.41	0.63	0.68	0.34	0.50	0.38	0.34	0.51	0.96	0.34	0.49	0.67	0.53	0.71	0.91	0.33	0.95
	TM	0.57	0.56	0.85	0.59	0.58	0.58	0.84	0.40	0.62	0.53	0.48	0.74	0.77	0.39	0.59	0.46	0.40	0.42	0.79	0.35	0.34	0.78	0.57	0.82	0.80	0.33	0.89
	AFINN	0.57	0.55	0.55	0.58	0.57	0.57	0.82	0.49	0.75	0.52	0.50	0.76	0.72	0.44	0.67	0.47	0.44	0.41	0.87	0.43	0.43	0.78	0.56	0.80	0.67	0.35	0.46
	BING	0.56	0.56	0.54	0.58	0.57	0.57	0.77	0.53	0.78	0.52	0.52	0.51	0.54	0.47	0.46	0.42	0.40	0.34	0.62	0.51	0.50	0.76	0.53	0.78	0.56	0.40	0.58
	Average	0.55	0.53	0.66	0.56	0.56	0.55	0.79	0.44	0.66	0.51	0.48	0.66	0.68	0.41	0.55	0.43	0.40	0.42	0.81	0.41	0.44	0.74	0.55	0.77	0.74	0.35	0.72

Note 5.142 The table illustrates the overall performance for all attempts in this study. The upper part of the table shows the results for the four different Amazon book review training datasets. The lower part shows the overall results of the lexicon training datasets. For each of the algorithms precision, recall and the F-score were calculated on a class level. Grey shaded algorithms have not produced good results; they failed to distribute the entities over the classes.

Table 5:78 illustrates the overall performance of the two different approaches (*Amazon* book reviews and the lexicon approach). It can be seen that the *SVM*, *MAXENT*, SLDA, *BOOSTING* and *RANDOM FOREST* algorithms show more or less the same behaviour over the two different tries. For *GLMENT* and BAGGING, no or less satisfactory results have been recorded. This was caused by the nature of the underlying training dataset. *NNET* and *TREE* remain weak on the second try.

Looking at the individual "average"[34] precision scores, we can see an improvement for most of the classifiers. *SVM*, SLDA and *BOOSTING* perform better when constructed with the help of the Amazon Book Reviews, based on the simple average measure. Yet, looking at the individual results, it can be seen that the lexicon approach of the articles produces higher individual values for *MAXENT* and the *RANDOM FOREST* approach.

Comparing the recall values, which state how many instances have been labelled correctly, an improvement can be seen. The *MAXENT* values have reached the highest values of more than 0.55 on average.

It can be summarized that, based on this first performance comparison, none of the two approaches clearly outperforms the other regarding the underlying training dataset. It also confirms the superiority of the *MAXENT* and *SVM* algorithms to perform the best, irrespective of the underlying training dataset.

The advantage of the lexical approach, as already mentioned, lies in the fact that it is possible to compare the quality of the final testing dataset with the training dataset labels.

[34] The average measure gives an indication of the improvement, and should only be seen as a simple comparison.

Table 5:79 - Performance analysis of the FT test dataset

Training Model	Class	SVM			MAXENT			GLMENT			SLDA			BAGGING			BOOSTING			RANDOM FOREST			NNET			TREE		
		Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score
NRC	1	0.96	0.99	0.97	0.97	0.97	0.97	0.94	1.00	0.97	0.95	0.97	0.96	0.94	1.00	0.97	0.94	0.99	0.96	0.95	1.00	0.97	0.98	0.95	0.96	0.94	1.00	0.97
	0	0.00	0.00		0.20	0.18	0.19		0.00		0.11	0.00	0.01	1.00	0.03	0.05	0.00	0.00		1.00	0.20	0.33		0.00			0.00	
	-1	0.64	0.50	0.56	0.48	0.55	0.51	0.24	0.01	0.01	0.33	0.29	0.30	0.77	0.11	0.18	0.20	0.06	0.09	1.00	0.25	0.40	0.35	0.70	0.47		0.00	
Overall		0.53	0.50		0.55	0.56	0.56		0.34		0.46	0.42	0.42	0.90	0.38	0.40	0.38	0.35		0.98	0.48	0.57		0.55			0.33	
TM	-1	0.94	0.98	0.96	0.96	0.95	0.96	0.89	0.99	0.94	0.91	0.96	0.93	0.88	0.99	0.93	0.88	0.92	0.90	0.88	1.00	0.94	0.96	0.94	0.95	0.86	1.00	0.92
	0	0.33	0.04	0.07	0.24	0.24	0.24		0.00		0.00	0.00		1.00	0.05	0.10	0.08	0.02	0.03	0.98	0.22	0.35		0.00			0.00	
	1	0.77	0.73	0.75	0.71	0.74	0.73	0.80	0.30	0.44	0.59	0.53	0.56	0.69	0.27	0.39	0.35	0.32	0.33	0.92	0.23	0.37	0.57	0.82	0.68		0.00	
Overall		0.68	0.58	0.59	0.64	0.64	0.64		0.43		0.50	0.49		0.86	0.44	0.47	0.44	0.42	0.42	0.93	0.48	0.55		0.59			0.33	
AFINN	1	0.90	0.96	0.93	0.92	0.93	0.92	0.85	0.98	0.91	0.87	0.94	0.90	0.84	0.96	0.90	0.87	0.62	0.72	0.83	0.99	0.90	0.94	0.84	0.88	0.76	0.99	0.86
	0	0.07	0.01	0.01	0.17	0.17	0.17		0.00		0.15	0.01	0.02	1.00	0.07	0.13	0.07	0.01	0.01	0.86	0.25	0.39		0.00			0.00	
	-1	0.81	0.75	0.78	0.77	0.72	0.75	0.83	0.51	0.63	0.71	0.60	0.65	0.75	0.49	0.59	0.33	0.70	0.45	0.92	0.39	0.55	0.56	0.87	0.68	0.66	0.06	0.11
Overall		0.59	0.57	0.57	0.62	0.61	0.61		0.49		0.57	0.52	0.52	0.86	0.51	0.54	0.42	0.44	0.39	0.87	0.54	0.61		0.57			0.35	
BING	-1	0.83	0.91	0.87	0.84	0.88	0.86	0.77	0.92	0.84	0.78	0.88	0.83	0.75	0.89	0.81	0.78	0.28	0.42	0.79	0.92	0.85	0.87	0.65	0.74	0.65	0.66	0.66
	0	0.18	0.01	0.02	0.23	0.14	0.17		0.00		0.22	0.02	0.04	0.84	0.10	0.18	0.00	0.00		0.89	0.23	0.36		0.00			0.00	
	1	0.78	0.81	0.80	0.77	0.78	0.78	0.77	0.69	0.73	0.74	0.71	0.72	0.72	0.62	0.67	0.39	0.89	0.54	0.80	0.69	0.74	0.53	0.88	0.66	0.44	0.52	0.48
Overall		0.60	0.58	0.56	0.61	0.60	0.60		0.53		0.58	0.54	0.53	0.77	0.54	0.55	0.39	0.39		0.82	0.61	0.65		0.51			0.39	

Note 5.143: The table shows the results of the 36 different algorithms (nine each) which have been trained on the four different lexicon training datasets. The test dataset (11,948 articles from the FT) has also been labelled with the lexicon approach so that the performance measures (precision, recall and the F-Score) have been calculated. For comparison reasons, an average for each classifier has been calculated. The grey shaded classifiers are either unable to produce significant results or showed in the training session a poor performance.

Table 5:79 illustrates the calculated final results for the 36 classifiers which have been trained on the four different lexicon approaches and applied on the FT test dataset. All classifiers have been applied to the testing dataset, despite the poor performance of some of them in the training period. It can be seen that *GLMENT*, *NNET* and *TREE* were unable to produce significant results for any of the tries since they either failed to distribute the results over the three classes or showed a tendency towards one category; this was anticipated since all three classifiers performed poorly in the training session.

The remaining classifiers were able to show good precision and recall values. As in previous tries, the *MAXENT* classifier was able to produce the most robust results, and for the *TM* trained classifier, it reached a recall value of 0.644, which is so far the highest value in this whole study. For three of the four different lexical training sets, the *MAXENT* classifier outperformed the other classifiers. Only for the *BING* trained classifiers did the *RANDOM FOREST* classifier have a higher recall (0.612) and precision value (0.824).

It can be summarized that the performance of the classifiers over the three different tests has slightly improved. The same classifiers (*GLMENT*, *NNET* and *TREE*) were unable to produce sufficient results in any of the tries. On the other hand, *MAXENT* seems superior in comparison. Yet, a recall value of more than 0.60 leaves room for improvement.

5.6.2.2 GRAPHICAL INTERPRETATION

Excluding the poor performers from the set of classifiers, I have again analysed them in a graphical way against the recession periods. The following four figures illustrate the classifiers, which have been trained with different lexicons. This graphical interpretation is similar to the earlier performed analysis in chapter 5.6.1.2. The grey shaded areas in the diagrams illustrate the recession period between 2008q1 to 2009q2, as well as two quarters with negative GDP growth in the U.K. in 2012q1 and 2012q3.[35]

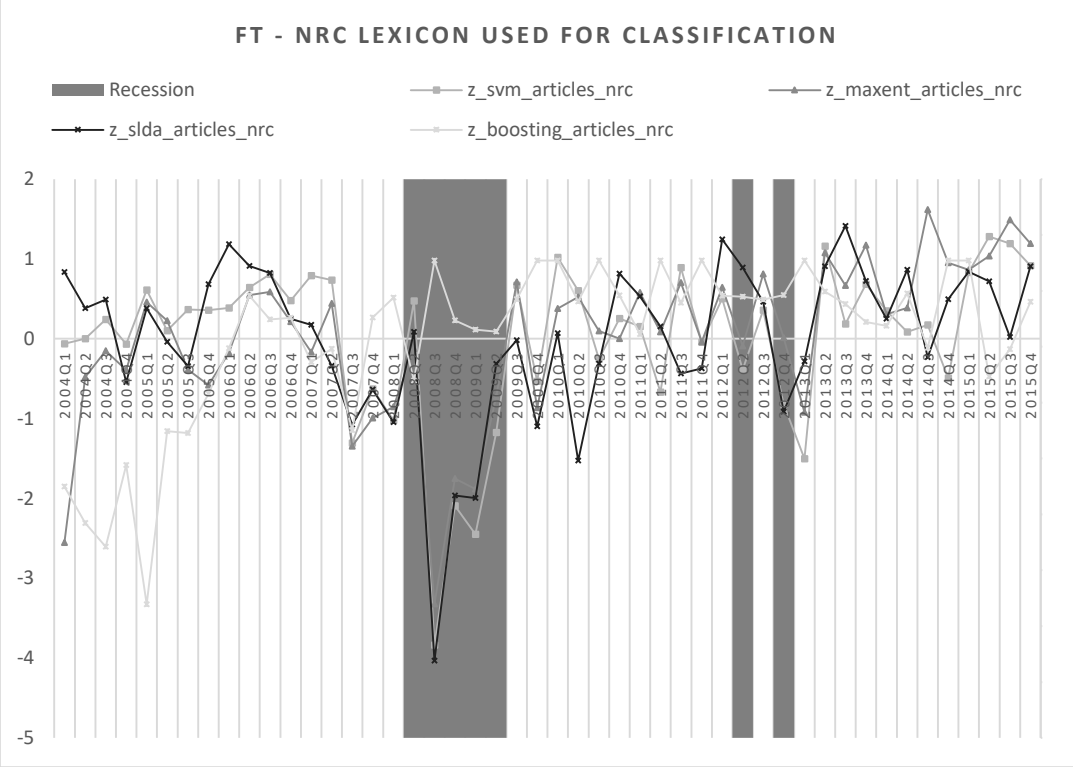
It is evident that in comparison to the above-shown results that the FT articles do not react as severely as the other sub-corpora. This has been discussed already in the previous chapters. The main reason for this might be the small number of articles per quarter.

For the classifiers trained with the *NRC* method (Figure 5:65), it can be seen that, after the values have been standardized, three of the four classifiers pick up the recession period in 2008.

[35] Data from the Office for National Statistics, https://www.ons.gov.uk/economy/grossdomesticproductgdp/timeseries/ihyq/qna, accessed on 14 December 2016.

However, they reach their lowest values in 2008q3 and start improving from there, while the recession continues for another three quarters. The *BOOSTING* classifier seems to react out of line and shows contradicting results to the other classifiers.

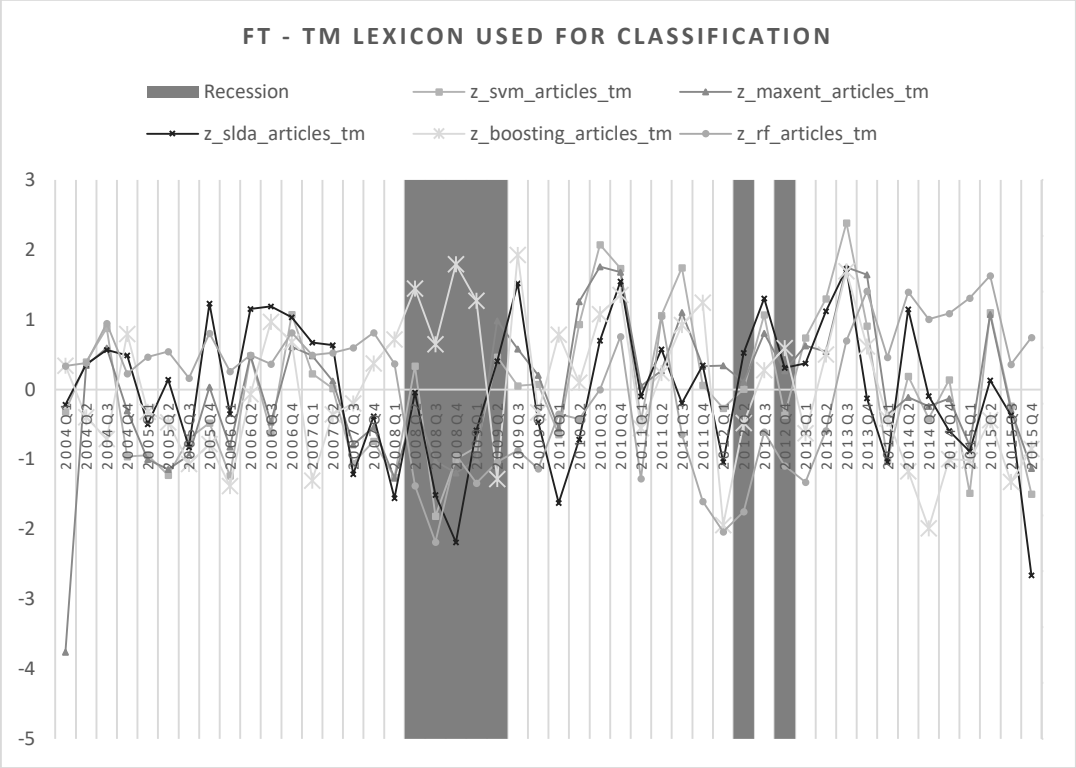
Figure 5:65 - NRC - Classifiers trained on an FT news corpus



Note 5.144: The figure illustrates the four classifiers trained with the FT news corpus - annotated by the NRC lexicon.

A similar result can be observed for these classifiers which have been trained with the lexicon used in the *TM* method (Figure 5:66). Out of the five classifiers, *BOOSTING* shows the same behaviour as before. On the other side, the classifiers seem to pick up the recession period. Yet again, their lowest point is more at the beginning of the period than at its end. It can be assumed that the textual sentiment indicators exceed the development in the market and that the reaction needs a couple of quarters to be reflected in the market itself.

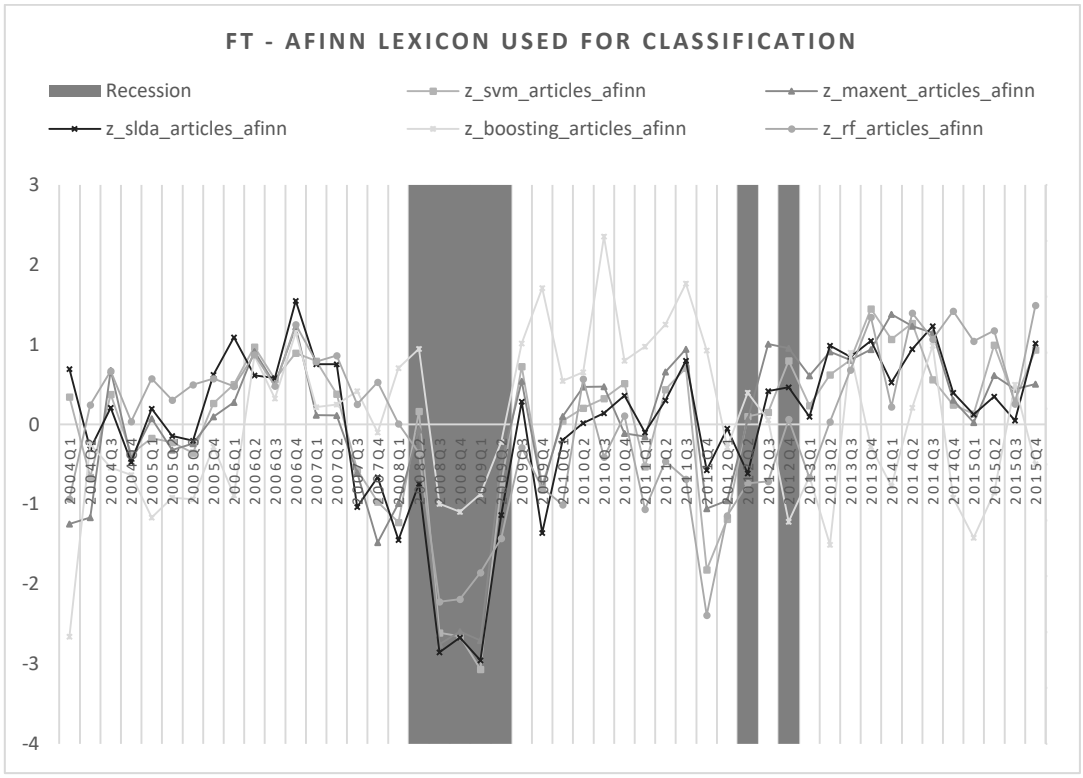
Figure 5:66 - TM - Classifiers trained on an FT news corpus



Note 5.145: The figure illustrates the four classifiers trained with the FT news corpus - annotated by the TM lexicon.

Figure 5:67 illustrates the indicators based on the *AFINN* lexicon approach. It can be seen that the indicators are much more in line with each other. One reason might be that the lexicon is based on the manual labelling of Finn (2011). Yet again, *BOOSTING* reacts much more severely than the other classifiers to changes in the underlying source. Besides, the classifiers precede the negative economic development in 2012q2 by two quarters, which has been established as the optimal lag for the textual indicators.

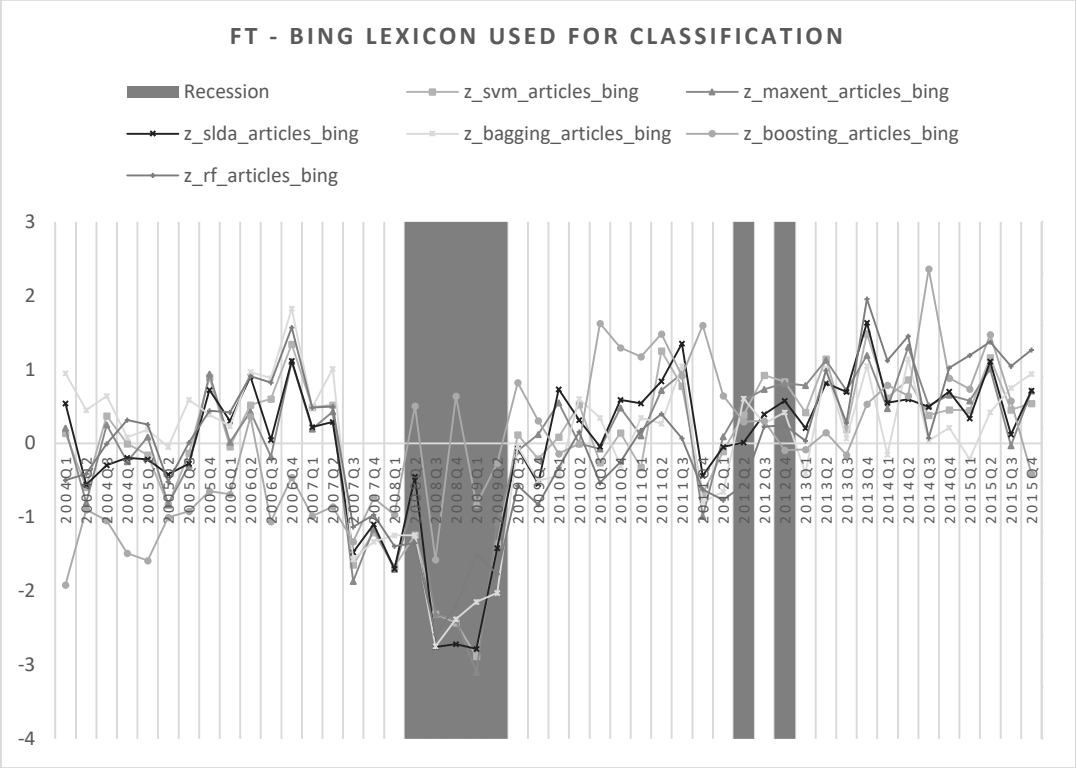
Figure 5:67 - AFINN - Classifiers trained on an FT news corpus



Note 5.146: The figure illustrates the four classifiers trained with the FT news corpus - annotated by the AFINN lexicon.

Finally, Figure 5:68 shows the result for the *BING* based classifiers. The results are similar to the *NRC* or *TM* results, with the *BOOSTING* classifier reacting oppositely to the other indicators. The reason for this behaviour in all four cases can be found in the weak performance measures. *BOOSTING*'s measures are by far the lowest in comparison and should be neglected here.

Figure 5:68 - BING - Classifiers trained on an FT news corpus



Note 5.147: The figure illustrates the four classifiers trained with the FT news corpus - annotated by the BING lexicon.

SUMMARY

The *AFINN* lexicon has produced the expected result where all four classifiers produce results in line with each other and in line with the trends caused by the recession periods, while the *NRC* results are reasonable, where the indicators react prior to the actual improvement within the market. Besides the *AFINN* model, the *BING* model has further produced good results, with the exception of the *BOOSTING* algorithm.

The following two tables illustrate the correlation between the newly constructed classifiers and the generated labels by the simple lexicon approach (Table 5:80) as well as the correlation to the originally constructed classifiers, which were trained with the *Amazon* book reviews (Table 5:81).

In the first table, all four combinations with the Maximum Entropy classifier, produce the highest correlation with the four different lexicons. This confirms the results of the above-shown values of the performance analysis. Table 5:81, on the other hand, does illustrate only a small share of combinations with a moderate correlation. This is somehow expected and surprising at

the same time. Expected, since the newly applied method is assumed to be more suitable given the weak results of the Amazon book review sentiment extraction. But surprisingly, in the sense that the initial sentiment values must have been partly wrong.

Table 5:80 - Correlation analysis - between new classifiers and labels from the lexicon approach

	*AFINN*_article	*BING*_article	*NRC*_article	*TM*_Net_article
SVM (AFINN)				
MAXENT (AFINN)	0.795			
SLDA (*AFINN*)	0.679			
BOOSTING (AFINN)	0.329			
RANDOM FORREST (*AFINN*)	0.527			
SVM (BING)		0.836		
MAXENT (BING)		0.852		
SLDA (*BING*)		0.841		
BAGGING (*BING*)		0.691		
BOOSTING (BING)		0.492		
RANDOM FORREST (*BING*)		0.802		
SVM (NRC)			0.404	
MAXENT (NRC)			0.706	
SLDA (*NRC*)			0.317	
BOOSTING (NRC)			0.443	
SVM (TM)				0.546
MAXENT (TM)				0.670
SLDA (*TM*)				0.261
BOOSTING (TM)				0.105
RANDOM FORREST (*TM*)				0.050

Note 5.148: The table shows the correlation between the labels from the newly created classifiers and the labels generated by the lexicon approach for the FT sub-corpora of the initially collected dataset for the full period 2004q1–2015q4. The left-hand column does further provide the total number of textual sentiment indicators generated by the combined method.

Table 5:81 – Correlation analysis - between the new and the original classifiers

	3c_eq_articles_*SVM*_old	5s_eq_articles_*SVM*_old	3c_eq_articles_max_old	3c_all_articles_max_old	5s_eq_articles_max_old	5s_all_articles_max_old	3c_eq_articles_SLDA_old	3c_all_articles_SLDA_old	5s_eq_articles_SLDA_old	5s_all_articles_SLDA_old	3c_eq_articles_BAGGING_old	3c_all_articles_BAGGING_old	5s_eq_articles_BAGGING_old	5s_all_articles_BAGGING_old	3c_eq_articles_*BOOSTING*_old	5s_eq_articles_*BOOSTING*_old	3c_eq_articles_rf_old	3c_all_articles_rf_old	5s_eq_articles_rf_old	5s_all_articles_rf_old
z_*SVM*_articles_*AFINN*_new	0.093	0.089																		
z_*SVM*_articles_*BING*_new	0.182	0.179																		
z_*SVM*_articles_*NRC*_new	-0.063	0.080																		
z_*SVM*_articles_tm_new	0.557	0.393																		
z_*MAXENT*_articles_*AFINN*_new			0.308	0.237	0.374	0.251														
z_*MAXENT*_articles_*BING*_new			0.227	0.282	0.233	0.229														
z_*MAXENT*_articles_*NRC*_new			0.417	0.290	0.394	0.291														
z_*MAXENT*_articles_tm_new			0.636	0.536	0.516	0.593														
z_SLDA_articles_*AFINN*_new							0.175	0.324	0.146	0.307										
z_SLDA_articles_*BING*_new							0.208	0.222	0.199	0.207										
z_SLDA_articles_*NRC*_new							0.234	0.156	0.157	0.259										
z_SLDA_articles_tm_new							0.336	0.559	0.238	0.424										
z_BAGGING_articles_*BING*_new											-0.038	0.033	-0.087	0.175						
z_*BOOSTING*_articles_*AFINN*_new															0.422	0.283				
z_*BOOSTING*_articles_*BING*_new															0.597	0.531				
z_*BOOSTING*_articles_*NRC*_new															0.514	0.564				
z_*BOOSTING*_articles_tm_new															0.197	0.227				
z_rf_articles_*AFINN*_new																	-0.136	0.197	0.153	0.144
z_rf_articles_*BING*_new																	0.078	0.187	0.298	0.186
z_rf_articles_tm_new																	-0.285	0.119	-0.018	0.257

Note 5.149:The table shows the correlation between the labels from the newly created classifiers and the labels generated by the original classifiers trained with the Amazon book reviews. The table shows only the correlation for the FT sub-corpora of the initially collected dataset for the full period 2004q1–2015q4 and compares the classifiers based on their methodology (i.e. SVM_new vs SVM_old).

5.6.2.3 FLEISS AND COHEN'S KAPPA

Besides the different comparisons in the above-described tests, it is further possible to analyse the similarity among the different newly constructed classifiers and the similarity between the lexicon labels and the supervised learning algorithms in a statistical way.

Different measures to compare the annotation of multiple annotators have been developed in the past. In the following, I am going to present the Fleiss kappa and Cohen's kappa measure. The Fleiss kappa measure, named after Joseph L. Fleiss, compares the agreement among multiple annotators in a classification task and belongs to the class of inter-rater reliability measures. The advantage over other measures, such as Cohen's kappa, is that multiple annotators can be compared at once. Fleiss (1971) defined the kappa as

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \qquad \textit{Equation 5:7}$$

where $\overline{P}$ is the observed actual agreement and $\overline{P}_e$ is the agreement achieved by chance. In the case where all raters agree, kappa takes a value of 1. Table 5:82 illustrates the possible interpretation of the kappa values.

Table 5:82 - Interpretation of Fleiss Kappa

Value of kappa	Interpretation
< 0	Poor agreement
0.01 – 0.20	Slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 1.00	Almost perfect agreement

Note 5.150: The table illustrated the interpretation of the possible Fleiss Kappa outcome.

Cohen's kappa is defined in a similar way. According to McHugh (2012), it is given by

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \qquad \textit{Equation 5:8}$$

where $\Pr(a)$ is the observed actual agreement and $\Pr(e)$ is the agreement achieved by chance. Table 5:83 shows the standard interpretation of the corresponding kappa values.

Table 5:83 - Interpretation of Cohen's kappa

Value of kappa	Level of agreement	% of data that are reliable
0–.20	None	0–4
.21–.39	Minimal	4–15
.40–.59	Weak	15–35
.60–.79	Moderate	35–63
.80–.90	Strong	64–81
Above.90	Almost perfect	82–100

Note 5.151: The table illustrated the interpretation of the possible Cohen's Kappa outcome.

In a first try, I have compared the inter-rater reliability of the nine newly constructed classifiers and the basic lexicon classifier. It can be seen in Table 5:84, that the *AFINN*, *BING* and the *TM* lexicon training datasets have led to a fair agreement among the classifiers. However, this first analysis also includes those classifiers which have been identified as poor performers. By removing them from the individual calculations, an improvement of the Fleiss kappa value can be achieved (Table 5:85), and the different *BING* classifiers even reach a moderate level.

Table 5:84 - Fleiss kappa for newly constructed classifiers - including all classifiers

	AFINN	BING	NRC	TM
subjects	11,948	11,948	11,948	11,948
raters	10	10	10	10
p-value	0.000	0.000	0.000	0.000
Fleiss kappa	0.370	0.380	0.214	0.351

Note 5.152:The table illustrates the inter-rater reliability among the different newly constructed classifiers. The analysis is performed only for the FT sub-corpora with 11,948 articles. The ten different classifiers for each of the four different underlying training datasets are the basic lexicon classification, SVM, MAXENT, SLDA, GLMENT, BOOSTING, BAGGING, RANDOM FOREST, Neural Network and TREE.

Table 5:85 - Fleiss kappa for newly constructed classifiers - without the poor performer

	AFINN	BING	NRC	TM
subjects	11,948	11,948	11,948	11,948
raters	6	7	5	6
p-value	0.000	0.000	0.000	0.000
Fleiss kappa	0.391	0.412	0.297	0.402

Note 5.153: The table illustrates the inter-rater reliability among the different newly constructed classifiers. The analysis is performed only for the FT sub-corpora with 11,948 articles. For the BING approach GLMENT, NEURAL NET and TREE have been excluded. For the AFINN and TM approach, the BAGGING classifier has been dropped. For the NRC approach, the RANDOM FOREST classifier was removed, due to its poor performance.

In a second try, I have compared the inter-rater reliability of the nine individual classifiers with the corresponding basic classifications of the lexicons (i.e. *AFINN* vs *SVM_AFINN*). Table 5:86 illustrates the results and shows that some classifiers have a moderate Cohen's kappa value. This indicates a satisfying level of similarity in the ratings. It further confirms that to some extent the inherent characteristics of the underlying training dataset have been carried over to final classification.

Table 5:86 – Cohen's Kappa for newly constructed classifiers and the basic lexicon classification

	SVM	MAXENT	GLMENT	SLDA	BAGGING	BOOSTING	RF	NNET	TREE
AFINN	0.666	0.637	0.510	0.525	0.473	0.227	0.463	0.566	0.068
BING	0.627	0.599	0.531	0.521	0.473	0.125	0.579	0.415	0.154
NRC	0.460	0.468	0.009	0.237	0.139	0.057	0.370	0.425	0.000
TM	0.658	0.643	0.335	0.446	0.303	0.222	0.333	0.610	0.000

Note 5.154: The table illustrates Cohen's kappa for each classifier, which has been trained on an annotated corpus with the help of a sentiment lexicon (e.g. AFINN approach). Only the corresponding lexicon and classifier were compared.

5.6.2.4 IMPLICATION INTO THE PROBIT MODEL

For the analysis of the newly constructed supervised learning indicators for the FT sub-corpus, I will again use the previous probit models. I have decided to use only the *AFINN* and the *BING* induced sentiment models since these are the two which have in the general analysis produced sufficient results.

From the newly constructed indicators, I am going to use the *SVM*, the *MAXENT*, the SLDA and the RANDOM FORREST models with their *AFINN* and *BING* versions. They will be compared to the lexicon-based classifiers *AFINN* and *BING*. For the dependent variable, I will use the converted *MSCI* capital growth rate for offices in London Mid-Town and West End. I have

decided to stick with these dependent variables since they have produced satisfying results. The testing period is between 2004m1 and 2015m12.

Table 5:87 illustrates the regression results for the newly constructed supervised learning sentiment algorithms. It can be seen that all ten indicators have a negative highly significant coefficient at the 1% level. Nearly all indicators enter the regression with one lag or more. Only the Random Forrest (*AFINN*) model has no lag. The number of lags has again been determined by the AIC.

The results for the pseudo-R-squared value are astonishing. The unchanged standardized lexicon methods, which have been superior throughout the entire analysis of this chapter, are now being outperformed by the newly constructed sentiment indicators. Again, the *BING* lexicon seems to be superior compared to the *AFINN* lexicon, since those learning algorithms based on the *BING* reach higher pseudo-R-squared values. The highest value is reached by the *SVM* (*BING*) model with 0.588. This value is not only more than twice as high as the original *BING* value (0.244), it is further the highest pseudo-R-squared value generated by any of the textual sentiment indicators. The indicator shows further statistically sufficient results, meaning that the Hosmer Lemeshow chi-square test is passed and that the classification score with 91.67 is based on a reasonable classification result.

Table 5:87 - Probit regression results for the newly constructed supervised learning algorithms

VARIABLES		(1) *AFINN* articles	(2) *BING* Articles	(3) SVM (AFINN)	(4) SVM (BING)	(5) MAXENT (AFINN)	(6) MAXENT (BING)	(7) SLDA (*AFINN*)	(8) SLDA (*BING*)	(9) RF (*AFINN*)	(10) RF (BING)
z_*AFINN*_article = L,	Standardized values for the lexicon approach with the *AFINN* lexicon	-0.607*** [0.144]									
z_*BING*_article = L,	Standardized values for the lexicon approach with the *BING* lexicon		-0.827*** [0.173]								
*SVM*_articles_*AFINN* = L,	Standardized values for the lexicon approach with the *SVM* (*AFINN*)			-0.827*** [0.157]							
*SVM*_articles_*BING* = L,	Standardized values for the lexicon approach with the *SVM* (*BING*)				-1.835*** [0.348]						
*MAXENT*_articles_*AFINN* = L,	Standardized values for the lexicon approach with the *MAXENT* (*AFINN*)					-0.729*** [0.141]					
*MAXENT*_articles_*BING* = L,	Standardized values for the lexicon approach with the *MAXENT* (*BING*)						-1.589*** [0.301]				
SLDA_articles_*AFINN* = L,	Standardized values for the lexicon approach with the SLDA (*AFINN*)							-1.191*** [0.225]			
SLDA_articles_*BING* = L,	Standardized values for the lexicon approach with the SLDA (*BING*)								-1.592*** [0.298]		
rf_articles_*AFINN*	Standardized values for the lexicon approach with the RF (*AFINN*)									-0.560*** [0.132]	
rf_articles_*BING* = L,	Standardized values for the lexicon approach with the RF (*BING*)										-1.206*** [0.223]
Constant		-1.163*** [0.149]	-1.271*** [0.166]	-1.274*** [0.166]	-1.962*** [0.318]	-1.195*** [0.156]	-1.796*** [0.275]	-1.494*** [0.210]	-1.745*** [0.261]	-1.072*** [0.140]	-1.577*** [0.229]
Observations		144	144	144	144	144.000	144.000	144	144	144	144
Log-likelihood		-53.02	-47.82	-44.77	-26.7	-48.910	-28.920	-37.21	-29.68	-56.39	-37.74
LR Chi2		20.45	30.85	36.96	76.35	31.940	68.660	52.07	67.13	20.16	51
Lag		2	2	2	1	1	2	2	2	0	2
Pseudo-R-squared		0.162	0.244	0.292	0.588	0.246	0.543	0.412	0.531	0.152	0.403
AIC		110.043	99.641	93.530	57.409	101.821	61.835	78.426	63.360	116.772	79.490
BIC		115.983	105.580	99.470	63.349	107.761	67.775	84.366	69.299	122.712	85.429
Correctly classified (%)		84.720	86.110	85.420	91.670	85.420	92.360	90.280	90.970	84.720	85.420
Sensitivity		8.700	26.090	26.090	66.670	25.000	60.870	52.170	60.870	220.000	39.130
Specificity		99.170	97.520	96.690	96.670	97.500	98.350	97.520	96.690	98.320	94.210
Hosmer-Lemeshow χ^2		7.990	18.120	3.970	2.610	4.560	0.870	7.030	1.270	1.870	3.440
Prob > χ^2		0.435	0.020	0.860	0.957	0.803	0.999	0.534	0.996	0.985	0.903
area under Receiver Operating Characteristic (*ROC*) curve		0.800	0.823	0.867	0.955	0.849	0.947	0.899	0.940	0.775	0.910

Standard errors in brackets (*** p<0.01, ** p<0.05, * p<0.1)

Note 5.155: The table above illustrates the probit regression results for the 10 selected newly constructed textual sentiment indicators. The dependent variable is the MSCI *capital growth rate for offices in London Mid-Town and West End. All indicators are highly significant at a 1% level and show the expected negative sign. The columns one and two apply the AFINN and the BING lexicon-based measures as they are. The columns three and four use the SVM indicator which has been trained by either the AFINN and the BING lexicon. The columns five and six use the MAXENT indicator which has been trained by either the AFINN and the BING lexicon. The columns seven and eight use the SLDA indicator which has been trained by either the AFINN and the BING lexicon. And finally, the last two columns utilize the RF measure. In all cases those indicators, which have been constructed with the help of the BING measure, are superior.*

The presented results suggest that the supervised learning algorithms based on the lexicon methods extract the sentiment incorporated in the articles. The indicators not only outperform the lexicon methods, but they also produce the highest results in this chapter. On the other hand, the result suggests that the *Amazon* book reviews are insufficient when it comes to the training of classifiers. My approach, to leave the provided code for the supervised learning algorithms untouched to allow for reproduction of my results, might have caused some of the insufficiencies in the above-presented analysis.

5.7 CONCLUSION

The detailed analysis of the various sentiment indicators has shown that sentiment can be extracted from news articles. The coverage of current events by significant newspapers provides enough data about the commercial real estate market. In the above analysis, I collected a unique dataset with more than 100,000 news articles for the commercial real estate market in the U.K. These articles have been classified in a two-folded way. First, I applied a lexicon-based approach, where the individual words of each article are classified into a specific category and there are then aggregated into a document specific score.

The second approach used nine different supervised learning algorithms to classify news articles. While the lexicon approach can be applied without any issues to any kind of document, the supervised learning approach requires a training dataset which is used to train the classifiers. The problem I faced was that there is no classified training dataset available. My initial idea to use *Amazon* Book reviews as a training dataset has been proven only partly suitable for the task at hand.

Various issues such as rating confusion (e.g. excellent was rated between three and five stars) and the unknown quality of the trained classifiers caused weak results in the subsequent modelling. A way around this could have been the manual labelling of the articles, by reading them myself or by another person or a group of persons. The problem with the first case is that

100,000 articles would take an enormous amount of time to classify. Second, my personal biases would influence the ratings I give. The same applies to the second possibility. Multiple raters would create the problem that not only one bias but various biases would irritate the process. Questions that could influence the manual rating of documents are: Is the person familiar with the real estate market? Has he had any bad experiences with the real estate market? A computer-based labelling process could overcome those issues.

My primary results suggest that the *Amazon* book reviews are unable to provide enough information in terms of training classifiers for the task at hand. Compared to the four lexicon approaches, the supervised learning algorithms were only partially able to improve the probit models. The lexicon approaches, invariably outperformed the supervised learning algorithms, in term of R-square values and sometimes even in terms of significance. The *BING* model especially has proven itself to be superior compared to any other classification method.

I have further shown in the four robustness checks that the classifiers are superior to the previously constructed sentiment measures. The advantage of the news articles as a source of sentiment is the frequency and nearly instant availability. In this study, I have transformed the extracted sentiment values into quarterly and monthly values, though I could have also used daily aggregations.

Compared to survey-based measures (e.g. the RICS sentiment survey), the newly constructed textual sentiment indicators did show high to moderate correlations but unfortunately failed to outperform the measures in a probit framework.

I have shown that a topic related training dataset is of vital importance to the classifiers. The ratings of the book reviews have been sometimes confusing, the wording of the reviews not bridging this issue sufficiently. Graphical analyses and the results of the probit regression have shown that sentiment can be extracted with *Amazon* book review ratings, yet not to the extent that a more straightforward and a less complicated measure could.

If the lexicon approach performs similarly to the supervised learning method, or even better, then the additional value for the use of more complex methods is questionable. Both the time and the complexity speak against their use.

Given these results, I was left wondering if the predictability of the supervised learning measures can be improved by combining the two methods. Therefore, I classified a training dataset with the help of the lexicon approaches. I then used the nine algorithms to train

classifiers based on this newly compiled dataset. The following created probit regressions produce outstanding results for the *Financial Times* sub-corpus.

The used training corpus for the test was a newly compiled dataset consisting of only *Financial Times* articles. To control for any seasonal sentiment swings (e.g. the financial crisis), the dataset is equally scattered over time as the test dataset. The reason for this is that any topic is not just influenced by the developments within the field at that time, but our feelings and actions are also influenced by our environment and other information we consume.

The constructed sentiment indicators are quite sensitive to the dependent variable in the probit model. In the above-presented results, I initially used the *MSCI* all property capital growth rate as well as the *MSCI* all offices capital growth rate. Based on the idea that a more targeted corpus should provide a purer market sentiment, I created five sub-corpora. However, the results of these tests were quite poor. The overall indicators have worked well for the two dependent variables, as well as the 100,000 sub-corpus results. Changing the dependent variable to a more London specific variable improved the results tremendously. One reason for that can be found in the weight of the London commercial real estate market within the country. Following the presented results, focusing on the largest and most read newspapers should provide sufficient insight into the market sentiment.

The shortcomings of the results are that the numbers in the articles are excluded by both approaches. This is a problem since we are dealing with economic topics in which numbers play a vital role for many people to judge market developments. There is a difference as to whether the market decreased or the market decreased by 50%. Here, a manual labelling exercise could help to bridge this issue.

While the goal of this chapter was to extend our knowledge and to test the practicability of more advanced sentiment measures, I have kept both the datasets and the code for the individual supervised learning algorithms untouched. Future work will include the extended analysis of the *SVM* approach including different kernel functions. A promising approach in this direction can be found in Kumar and Gopal (2008) who developed different approaches around the *SVM*. Further, could a better dataset improve the results of the classifiers? The general search of news articles is very likely to incorporate no real estate related entities.

GLMENT and other algorithms allow for further fine tuning. It seems promising to investigate well-functioned algorithms even further. The applied methods could also be transferred to other regression-based analysis in the real estate field. I have tried to show the

advantage of the classifiers by applying them to the probit and a standard yield model. Future work will include a much more detailed and customized application of these indicators.

Since most of the classifiers are initially developed for binary classifications, it might be suitable to increase the performance by dropping the neutral entities and only focus on the positive and negative observations in the training dataset. This could produce better results for the neural net and the decision *TREE* approaches.

In this trial and within the literature the classifiers remain on the small side of the training corpus, due to the 20% - 80% split. It might improve the results when the classifier is retrained after it has been identified as a good performer. Then the classifier would rely on 100% of the training data, which would add further information.

Another improvement of the results could be achieved by reusing the statistical modification method from chapter 3.4.2. Orthogonalizing the textual sentiment indicators against observable facts could lead to a purer market sentiment.

To conclude, the *BING* method, as well as a focus on the mainstream newspapers, could provide market participants with enough insight into market development.

6 CONCLUSION

6.1 AN OVERVIEW OF THE THESIS

Following the definition of Baker and Wurgler (2007), the sentiment is the belief of market participants about future cash flows and the investment risk that is not justified by the facts at hand. In other words, sentiment can provide an aggregated measure of the opinions and the beliefs of market participants about future developments. Motivated by the observation that investors do not act as rationally as assumed, sentiment analysis has been used to provide an idea of their irrational behaviour. Studies such as Carroll et al. (1994); Baker and Wurgler (2007); Clayton et al (2009); Tsolacos (2012); Dietzel et al. (2014); Marcato and Nanda (2016); Freybote (2016) or Heinig and Nanda (2018) have shown that sentiment plays a vital role in equity and real estate markets.

The majority of real estate studies have focused on the US housing market. The European commercial real estate market has been largely excluded from sentiment analysis. The reasons for this avoidance can be found in the fact that the housing markets are subject to more transactions and therefore to better and more rapid absorption of sentiment swings. Further, analysis of the US market allows a higher degree of comparability when it comes to economic and real estate specific measures across different regions and cities.

However, the European commercial real estate market is one of the largest investment markets in the world and is also subject to sentiment swings. Therefore, a sentiment analysis, given the knowledge that investment decisions are seldom performed in a rational framework, should be performed.

The second motivation which has driven this thesis is the absence of a universal sentiment proxy. While some markets do have a direct sentiment measure, such as the U.K., many other countries don't. This makes it somewhat difficult for investors and scholars to extract the underlying belief of market participants. Even where a direct measure exists, it might not be comparable to those of other countries due to differences in structure. Therefore, indirect sentiment measures are used. Some scholars such as Ling et al. (2014) use REIT related measures to extract the market specific sentiment. However, these approaches require the existence of a functioning REIT market within the countries of interest. In the first study of this thesis, I used a range of different European countries, including East European countries that

do not have similar market structures and where construction of sentiment measures based on REIT indicators is impossible. Other approaches have utilized only one measure at a time, such as the architectural billings index [Baker and Saltes (2005)]. These approaches are one-sided and exclude the wider picture of the market.

Using multiple sentiment proxies requires statistical modification. However, most of these proxies initially measure other things in the first place. This leaves room for doubt as to whether the extracted sentiment does equal the actual sentiment of the market. Further, the publication time of these proxies is very important. Depending on the proxies used, this could be up to three months behind the actual observation. Therefore, only an *ex-post* analysis is possible.

This has driven the search for an updated measure which is closer to the market. One suitable approach is the use of online search volume queries, which allows drawing on the thoughts of millions of people. Tools such as Google Trends have massively improved forecast models. One could argue that the use of online search volume indicators does not initially provide a suitable sentiment indicator since search queries only provide searches of interest and not actual actions. However, the main advantage of the tool lies in the fact that it is available and comparable for and between different markets.

Approaching the topic of sentiment should, therefore, start with the question of how we make our decisions. Three possible areas that contribute to our decision-making process have been identified: discussions with friends and colleagues, personal experiences and newly acquired information. The last part can be measured in a scientific framework. Most of our information is stored in texts. This allows the extraction of the sentiment from these text documents.

The idea behind the utilization of texts as a proxy comes from the fact that we all read to broaden our minds. In an investment case, where we do not know anything about a new market, we require information. This can be either included in market reports, where service agencies provide an aggregated view on the specific market, or they can be included in newspaper articles. The latter group is more likely to provide a general description of the market but has a higher frequency when it comes to publication.

The three presented empirical studies of this thesis have been produced in accordance with these thoughts. Before I will describe in detail, which specific contribution was made by each chapter I like to summarize them more generally. The contribution to the literature is that I have shown that European real estate markets are subject to sentiment swings on a large scale. The

use of various sentiment proxies for different countries makes it possible to compare markets with each other. That has been impossible so far, since chosen sentiment measures where market or data specific. The second major contribution of the thesis is that other mediums such as text documents allow us to extract sentiment. Newly developed methods allow an easy and straightforward application of sentiment extraction. Changing the methodology and using a universal information source allows not only to compare markets with each other, but it does also allow to get an updated sentiment measure at any time. While two methods have been tested it was shown, that the combination of both word lists and supervised learning algorithms produce the best results.

In Chapter 3, I focused on the European commercial real estate market. A large dataset of 24 European countries with 80 city regions was analysed. The dataset represents a mixture of different countries that are in different stages of their market development. City regions located in the Western European countries are characterized by a higher degree of transparency and liquidity. In general, more information about the different real estate sectors is available which allows investors to make sound decisions. Eastern European countries, on the other hand, show a different stage of real estate market development, where national and international investors only slowly enter those markets. Poland, for instance, is a good example given its recent developments over the last decades. Another sign of the current stage of the market is the existence of various service providers. The more market players are present, the higher the degree of transparency and information. However, mainly this scarcity of information allows sentiment to play a more vital role in the real estate markets.

The structure of Europe with the European Union and the Eurozone makes it challenging to find an overall indicator which is published continuously and applicable for all countries within the dataset. Direct sentiment indicators such as the published survey of RICS do not cover all countries. The Economic Sentiment Indicator published by the European Union, on the other hand, has the problem of excluding various countries and that it mostly deals with topics that are not linked to real estate. This makes it necessary to use sentiment proxies to generate an overall market indicator. In the first study, I decided to use a set of four primary indicators.

The macroeconomic indicator is constructed with the recommended method of Baker and Wurgler (2007) – a combination of an orthogonalization process and a PCA. I used six different sentiment proxies, which were regressed against observed macroeconomic variables. As sentiment proxies, I have used two direct measures the BSI and the ESI, both published by the European Union. In addition, four indirect measures were applied: the change of the stock

market, change of consumer confidence, the national credit rating as well as the 10-year government bond rate. Those factors were widely available for the countries within the study and in combination, they provide a full picture of the economy of each country.

The second and third sentiment indicators use real estate specific variables in the orthogonalization process. As a sentiment proxy, I used the IPD total return series for both the office and retail market. During the orthogonalization process, I encountered further issues, due to data availability on the retail side. While the office sentiment proxy, was regressed against several observable factors, the retail proxy was only reduced by the market rent observations. Since only one proxy was used, a PCA was obsolete.

The last sentiment indicator was developed by the motivation that online search volume indicators provide a sufficient amount of information about the markets. I used Google Trends to extract city region-wide search volume scores for 90 different search words. The aggregation of these scores generated an individual online search volume indicator per city region. Online search volume measures have become widely accepted and a large body of literature is now developed. The idea is it to proxy the interest of market players at an initial stage when people start gathering information. However, online search volume indicators do not guarantee that an actual market action took place.

Those four indicators were then introduced to a standard yield model. My results have shown that adding any of the indicators causes the resulting model to outperform the base model. For the office market, the online search volume measure reached the highest pseudo-goodness of fit score with 0.852, compared to the base model with 0.826. The office specific indicator ranks second, followed by the macroeconomic index.

For the retail market, this picture is slightly different. All three indicators still outperform the base model; however, the macroeconomic measure produced the highest value with 0.791. The retail market specific measure came second, and the online search volume index only produced slightly better results than the base model.

Further tests have shown, that sentiment induced yield models to perform better when it comes to forecasting estimations. However, these results differ from city region to city region.

I extended the study by analysing further possible combinations of proxies and methods. For instance, in Baker and Wurgler (2007) the PCA relies solely on the first principal component. Different approaches are possible, for instance using all components with an eigenvector larger than one. However, by switching to the Kaiser Criterion the results have remained more or less

similar and, given the more complex way of constructing the measures the initial proposed way should be favoured. Yet, the combination of both methods, the orthogonalization and the PCA, with a focus on the first principal component is superior in comparison to other methods.

I have further analysed whether the produced results might have been strongly influenced by the composition of the dataset. The German, French and British markets carry a larger share in the dataset. I have, therefore, split the dataset into two shares: One including these three markets and the second set with all the remaining city-regions. The results suggest that both market shares rely on different sets of sentiment measures. While the more established markets did reveal a stronger tendency to the property specific indicators, the remaining city-regions did rely on the macroeconomic and online search volume measure. This suggests that property specific information is probably less reliable and that market participant make their decisions preferably with the help of general market information. The better result for the online search volume measure, on the other hand, can be argued for with the same logic. Due to the absence of prominent market players, which in general provide more market transparency, more excessive information gathering is performed online. Therefore, the online search index produced better results.

In addition, this finding has allowed me, to compare more general, the underlying study to the equity and fund market. Mian and Sankaraguruswamy (2012) or Lee et al. (1990) have analysed the closed-end fund puzzle. Here, and that is similar to my finding, small, young, highly volatile and non-dividend paying stocks are more exposed to sentiment shifts. That again is caused by the lack of transparency and information scarcity.

Nevertheless, the question remained as to whether the constructed sentiment indicators do actually measure the sentiment of the market? Also, the construction process can be described as complex and time-consuming. Yet, the strongest concern against the use of macroeconomic sentiment measures arises given the different time frames. When the sentiment proxies are published, the market has already moved on and the provided signal might be already outdated due to new developments.

Motivated by these observations, I focused on the U.K. commercial property market in the second study. Similar to Soo (2015) and Walker (2014a, 2014b) I identified text documents as a promising source for the extraction of sentiment. Since the U.K. market is one of the major real estate investment hubs in Europe, a variety of service agencies are present. One of their main marketing tools is the publication of market reports. These reports represent a summary of the

most recent market developments and provide an outlook as to what market participants might expect.

Word lists allow the classification of documents into either a positive or negative category. Through the aggregation of multiple documents per quarter, market and property class, specific sentiment scores were developed. The results of this second study revealed that market reports carry market sentiment. Autoregressive models, that have been induced with textual sentiment indicators produce higher R-squared values. From the three presented panels in the second study, those which are more focused on a specific market segment produced much better results. The office market reports related to London gave a better indication for the estimation of the IPD total return index.

While this first application produced sufficient results, even in comparison to the previously applied sentiment indicators, some drawbacks were observed. First, none of the four textual sentiment indicators produced superior results in comparison to the other three. Only the *NRC* lexicon was identified as the weakest among them. One reason could be the original background of each of the four sentiment lexica. Given the fact that the NRC dictionary was originally developed to extract emotions rather than sentiment, the poor result in this second study and later on seem reasonable. Both the BING and the AFINN lexica produced rather robust results. The Topic Modelling (TM) method, based on the Harvard General Inquirer Dictionary, showed the most promising results.

The performed correlation analysis between the direct (RICS) and indirect (textual sentiment indicators) measures only produced weak to moderate results. This leaves room for doubt about the quality of the newly constructed indicators. Finally, and this represents the main problem of the second study, the number of documents, which were used for the construction, is rather small. The textual sentiment indicators are based on 150 to 819 market reports spread over up to 35 quarters. In addition, the total number of reports used per quarter is smaller at the beginning of the testing period than towards the end. Therefore, those sentiment scores are only based on a few documents, which makes them much more judgemental.

However, I assume that the underlying medium for the sentiment extraction is better suited than the macroeconomic sentiment proxies. Market reports are much more linked and focused towards the market and they should allow a better and closer look on the current developments. Another advantage is the possibility to focus on specific asset classes within specific regions. Given the moderate results of the second chapter, I come to the conclusion

that the market which is going to be examined and the corresponding sentiment should be linked. This has allowed to structure the third and last analysis of my thesis accordingly.

I decided to use a more robust dataset which could support my hypothesis that text documents carry the market sentiment. Newspaper articles were used in other studies and provide a source of information on a daily basis. The third and most extensive empirical study of this thesis tried to tackle the previously encountered issues. I not only applied a different dataset, but also used a more advanced method to extract the sentiment from text documents.

Supervised learning algorithms have been applied in various other disciplines. I order to test which method could extract the sentiment better, I compared nine different methods. All methods essentially share a similar approach and some are extensions of others. Other methods such as the neural network are rather complex in the way how the sentiment measure is formed. In general, all methods require two datasets that are similar in their underlying structure. The training dataset, where various text documents have already been labelled, and a test dataset with no labels. Unfortunately, no labelled document corpus is available for the U.K. and the real estate market. My initial idea to bridge this circumstance by using *Amazon* book reviews only produced weak results. The idea was, to use book reviews, that have been given to real estate related books. I assumed that these books are read by professionals or soon to be professionals. And given that, I hoped by covering multiple real estate topics to generate a large enough training corpus, which essentially should have been similar to the text in the news articles.

One reason for the poor performance of the applied method could be the provided ratings of the book reviews. They were in multiple cases rather diverse and inconsistent (Table 5:3). In addition, the book reviews seem to differ in their wording compared to the newspaper articles. The method still produced sentiment indices, but compared to the earlier introduced lexicon approaches, these were rather weak in their performance. The results of this study favour the four different lexicon approaches, and especially the *BING* and the *AFINN* methods.

While these problems were easily traced back to the very nature of the training dataset, another set of issues arose out of the applied methodology. I realized that the sorting task for most of the algorithms were either too complicated or unsolvable at all. Sorting entities into one of 5 different categories minimizes the nuances between these categories and makes a final decision more difficult. This has been observed by the fact that some algorithms sorted the entities entirely into one category and ignored the other. A second issue was that the collected book reviews dominated by positive ratings. An algorithm trained on these would, therefore,

be more likely to sort the news articles into one of these classes. I have decided to deal with both issues by applying different approaches. I constructed sentiment measures based on the full book review data set and on an equalized dataset. By using an equalized approach, I lost more than 80% of my observations. As described in section 5.3.2 the lowest share of collected reviews had a total number of 7,548. In order to construct an equalized corpus, I reduced the number of observations in each category down to this number. This could have caused more suitable reviews to be rejected. In the other case, I was forced to limit the number of categories to three, by assuming, that a given rating of three stars would mean a neutral categorization of the book. The classes one and two where than combined to the negative group and four and five to the positive group. By using the full review corpus, the tendency to the positive class was still given.

The test dataset used a total number of 109,103 collected news articles. Due to the observation in the second study, that the sentiment indicators perform much better when the sentiment is extracted from a targeted source, I have sliced the full corpus into five sub-corpora. Each corpus was selected, with the motivation that the underlying articles shared a similar structure or content, and that sentiment extracted from these articles was either more directed or more suitable for the prediction of the dependent variable.

In a first analysis, I decided to analyse each set of indicators for each of the five different sub-corpora in a graphical way and plotted them against the recession period of the U.K. I wanted to verify, if there is a common trend among the different methods and towards the general economy. After this simple analysis, I decided to remove those indicators failed to extract a comparable sentiment. This has been done in accordance to the performance analysis. Here the algorithms are tested against a retained share of the of the labelled observation. As pointed out earlier, the lexicon approach indicators produced extremely good results, and out-performed all supervised machine learning measures in all tries. In total, eight indicators entered the probit models of each sub-corpora.

The first corpus used all collected news articles. While I assumed that this corpus was very likely to carry noise, the performance of the indicators based on the full article set were superior in comparison to the other sub-corpora. A reason for this rather surprising result can be found in the fact that I initially used two broader dependent variables. On the other hand, this result confirmed my initial hypothesis that more general news adds to the market sentiment. Arguably, the more specific an information or data source is, the less likely is it that the information will impact the general market sentiment. This seems reasonable, since the asset is

not traded in an isolated vacuum, but in a complex market structure. Using a broader information hemisphere allows for important topics to gain momentum and impact in the market. Multiple opinions towards one topic will increase the awareness of news consumers regarding the issue and might lead to an adjustment of behaviour.

The second sub-corpus has been constructed with a smaller dataset, which excluded all those articles having housing related words. The intention was to reduce the noise of the corpus and to construct a more focused set of documents regarding the commercial real estate market. The initially used search words when I collected the articles were focused on the commercial real estate market, but many articles discuss two or more asset classes at once. So, housing related topics where accidently collected. Different to my initial assumption, the removal of housing related articles, did not increase the results. The supervised learning measures suffered an essential loss in their significance. Only the two Maximum Entropy models performed reasonably well.

The third corpus was designed to provide a focused view of the London market, and only news articles which included the word “London” were considered for the construction of the sentiment indicators. Here again, the already observed pattern did continue and the same indicators dominated the probit results.

The two remaining corpora did not directly try to change the focus of the underlying sentiment, but to readjust the main source of information. While the full set of articles included a range of various small newspapers, I assumed that newspapers with a broader coverage should carry a more severe sentiment. Finally, the last corpus tried to apply this idea in a more extreme trial. I only considered *Financial Times* articles for the construction of the sentiment indicators, with the motivation that the newspaper is very likely to be read by real estate market participants.

To summarize, the results of the four different sub-corpora were unable to produce more satisfactory results compared to the overall corpus. This was not only true for the supervised learning, but also for the lexicon approaches. However, this picture changed, when I changed the underlying dependent variable. The two MSCI series used in the initial try were broad market measures and were not focused enough. Therefore, I switched the underlying dependent variable once more. I introduced two London specific *MSCI* capital growth rates. An improvement in the performance of the sentiment indicators was observed. This proved my assumption that the sentiment within the articles extracted from a more focused sub-corpus should perform much better than an overall corpus. From the three chosen indicators, namely

the *AFINN*, the *BING* and the *MAXENT* I, the London and the 100,000 sub-corpora outperformed the remaining indicators. This is a satisfactory finding and underlines the earlier observation that both the sentiment and the dependent variable should share a common theme.

In a second smaller robustness check, I compared the performance of the textual sentiment indicators to the direct sentiment measures of the RICS. At least the BING measure was able to outperform the direct measures.

I have further tested the robustness of the newly constructed sentiment measures against all other constructed sentiment measures within this thesis. The flexibility of the news measures to change the aggregation from monthly to quarterly does allow these comparisons. I have applied the textual sentiment measures to the standard yield model from chapter 3. The best BING model from the last chapter has also here produced the best result according to the R-squared value.

Given the poor results of the *Amazon* book reviews and that they essentially have failed to provide a sufficient training dataset, I decided to extend the analysis. To revaluate the performance of the supervised learning algorithms, I tried to combine the two methods used in chapter 4 and 5. I collected another 55,872 articles from the *Financial Times* as a training dataset. Since these articles still miss the corresponding labels, I applied the four different word lexica to this corpus. Since the lexicon approaches performed reasonably well throughout the last two chapters, I assumed that the provided labels could generate a sufficient training dataset. Since the earlier results have been improved, by using only three categories, I decided to follow this method as well. This training dataset was then introduced to the supervised learning algorithms. The sentiment was extracted from the already existing FT sub-corpus. I performed another analysis in both a graphical way and in a statistical way. The improvement of the results was surprising. Especially, those textual sentiment indicators which have been trained by the BING lexicon have produced good results. The probit model for the Support Vector Machine indicator has produced a pseudo-R-square value of more than 0.588, for the model using the *MSCI* capital growth rate for offices in London Mid-Town and West End.

This last analysis produced enough robust results to prove the hypothesis of this chapter and this thesis. Real estate markets are subject to sentiment swings; however, the measurement of sentiment is sensitive to both the sentiment proxy used and the targeted subject. More focused dependent variables on both sides improve the results significantly.

To conclude, the method provided by Hu and Liu (2005) as well as Liu et al. (2005) generated the most robust results within the analysis undertaken in this thesis. The classification of text documents produced more reasonable results, when the training dataset was equalized and when the number of possible classes was reduced to three. Further, it seems that the Maximum Entropy algorithm, which tries to reduce the uncertainty of a dataset, is more suitable when it comes to the extraction of sentiment. However, I would like to point out one more time that the application of the different algorithms was performed without any modification of the code. Readjustment could have produced much more reasonable results.

Given the evidence in this thesis suggesting that market participants are influenced by external factors, such as news articles, the consideration of textual sentiment can moderate irrationality in the market. This means, that if we know about this circumstance and if the sentiment can be measured, we could act accordingly. And that would give the irrational element of the market a rather rational component, which could be exploit by businesses and other market participants.

6.2 LIMITATIONS AND FUTURE WORK

Sentiment analysis has become a significant field of interest. Various studies have found that real estate market participants are subject to sentiment swings. This has either been proven by the application of different sentiment proxies, or it was argued that the market is subject to sentiment due to the weaknesses of its characteristics. Since not all markets and not all property sectors are covered by direct sentiment measures, market participants need indirect sentiment proxies.

As I have shown, different kinds of proxies are available. However, mature and immature markets lack the existence of a universal sentiment proxy. The extraction of the market sentiment from newspaper articles has been found to be a sufficient information source. However, the results presented here just line up with the results from Soo (2014) and Walker (2014 a, b; 2016) and much more analysis needs to be performed.

During my work, I encountered various limitations, which have partly caused some results to remain weak or even questionable. In the first study, I encountered various data availability issues. Besides the fact that some macroeconomic variables were selected for some city regions from different data sources, the main limitation can be found with regards to the retail

sentiment specific indicator. While its office counterpart included six different observable market factors in the orthogonalization process, the retail measure only removed the rent variable from the IPD total return index, which I used as a sentiment proxy. Therefore, the retail sentiment indicator resembled much more strongly the original proxy and not so much the unexplainable element, which is likely to be included in the indicator.

Further, I would have liked to extend the work on the online search volume measure. Throughout the last years, the tool has been used in various studies as a sentiment proxy. The newest application of Google Trends allows for weekly and even daily downloads of the search interests. I could have used a monthly composite of online search volume to compare the results of the later studies in much more depth. Analysing the text documents, by topic modelling techniques could have also revealed topics and terms of interest within the market, which I could have used to generate an updated online search volume measure.

The second study has essentially two limitations. First, as has become clear, the number of market reports, which have been used for the construction of the different indicators, is too small. Not only is the number for the office specific measure only based on 150 reports, but they are also spread unequally over 35 quarters. This has produced a measure with more weight of the reports towards the end than at the beginning where the number of reports was lower. The market reports are published by the different service agencies and made publicly available on their websites. A sufficient number of reports could have been generated by getting in contact with the service agencies, or by constantly downloading those documents over a longer period.

The second limitation also occurred during the third empirical study. The standard way of pre-processing the different text documents excluded the numbers from them. For the word lexicon approach, this step is entirely understandable; however, for the supervised learning algorithms, a trial utilizing the numbers in the documents could have produced slightly different results. Since the topic is embedded in an economic framework, numbers play an essential role in the judgement of the information presented in the reports or even in the news articles. Unfortunately, I was unable to find in the literature any example where numbers were considered during the sentiment analysis. I assume that future developments and updated algorithms will incorporate numbers and the chance to estimate their meaning within text documents.

The low processing power of the computers used restricted a more complete calculation of all supervised learning algorithms. For the purpose of this investigation, I could have reduced

the number of articles in the training process. Since the majority of the nine applied algorithms produced results in all four training sets, I did not change the number of articles.

Another possible limitation could have been caused by the way of constructing the different algorithms. In hindsight especially, the equalized corpus with three categories should have constructed in a different way. While the general idea of removing a tendency towards any category was followed by using an equalized corpus, this has been, unfortunately, violated by combining the first two and last two categories. Therefore, the algorithms did have a stronger tendency to the positive and negative class but not to the neutral one. I should have either reduced the number of the categories in each of the classes, ignored the second and fourth class at all, or I could have increased the number of reviews in the third category since more observations were available. By considering this different angle, I could have produced more robust results for the equalized corpus.

Overall, I would like to extend my research in the future. In particular, I hope to improve the predictability of the various applied supervised learning algorithms. Since all of them allow for further modification during the process of construction, I should be able to generate more robust results when a modified and probably more flexible code is applied. Especially, the weak results of the Neural Network algorithms, have been surprising. However, due to the fact that the code for the training and testing step hasn't been modified, the result is maybe not that surprising. The Neural Network algorithm has become popular within the last years, since, in comparison, it does produce more robust results.

An important part of the construction of the supervised learning algorithms is the training dataset. As different research has shown, the existence of a labelled training dataset is essential to the process. It has further become clear that those datasets which are labelled by a human being are much more precise when it comes to the training of the algorithms. Therefore, one possible area of research could be the development of a labelled training dataset for the real estate market.

Future research will also include the extension of the work to other markets such as the German or French market. A multinational comparison study should allow the generalization of my findings and to take the research on sentiment analysis with the help of text documents a step further. For the German market, I am already in contact with a major information provider, regarding a new real estate related news article dataset. One goal of this market extension should be the automatization of the analysis process. I hope to generate via an API a daily or instantaneously updated news-sentiment-indicator.

I am also interested in extending the work regarding the direct sentiment measures. I found it somewhat surprising that not all countries have a similar direct sentiment market survey. I am aware of the problems, which I have pointed out multiple times in this study, but for market comparison reasons, an international sentiment survey would be beneficial for all market participants.

7 REFERENCES

Abbasi, A., Chen, H., & Salem, A. (2008). Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. ACM Transactions on Information Systems (TOIS), 26 (3), pp. 12:2 – 12:34.

Ahern, K. R., & Sosyura, D. (2014). Who writes the news? Corporate press releases during merger negotiations. The Journal of Finance, 69(1), pp. 241-291.

Aissia, D. B. (2016). Home and foreign investor sentiment and the stock returns. The Quarterly Review of Economics and Finance, 59, pp. 71-77.

Akins, B. K., Ng, J., & Verdi, R. S. (2011). Investor competition over information and the pricing of information asymmetry. The Accounting Review, 87(1), pp. 35-58.

Augustyniak, L., Kajdanowicz, T., Szymański, P., Tuligłowicz, W., Kazienko, P., Alhajj, R., & Szymanski, B. (2014, August). Simpler is better? Lexicon-based ensemble sentiment classification beats supervised methods. In Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on (pp. 924-929). IEEE.

Bai, X. (2011). Predicting consumer sentiments from online text. Decision Support Systems, 50(4), pp. 732-742.

Baker, K., & Saltes, D. (2005). Architecture billings as a leading indicator of construction. Business Economics, 40(4), pp. 67-73.

Baker, M., & Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. The Journal of Finance, 61(4), pp. 1645-1680.

Baker, M., & Wurgler, J. (2007). Investor sentiment in the stock market. The Journal of Economic Perspectives, 21(2), pp. 129-151.

Barber, B. M., & Odean, T. (1999). The courage of misguided convictions. Financial Analysts Journal, pp. 41-55.

Barber, B. M., & Odean, T. (2007). All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. The Review of Financial Studies, 21(2), 785-818.

Barber, B. M., Odean, T., & Zhu, N. (2009). Do retail trades move markets?. Review of Financial Studies, 22(1), pp. 151-186.

Barberis, N., Shleifer, A., & Vishny, R. (1998). A model of investor sentiment. Journal of financial economics, 49(3), pp. 307-343.

Barkham, R., & Ward, C. (1999). Investor sentiment and noise traders: Discount to net asset value in listed property companies in the UK. Journal of Real Estate Research, 18(2), pp. 291-312.

Beracha, E., & Skiba, H. (2011). Momentum in residential real estate. The Journal of Real Estate Finance and Economics, 43(3), pp. 299-320.

Beracha, E., & Wintoki, M. B. (2013). Forecasting residential real estate price changes from online search activity. Journal of Real Estate Research, 35(3), pp. 283-312.

Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. "O'Reilly Media, Inc.".

Black, N. T., & Ertel, W. (2011). Introduction to artificial intelligence. Springer Science & Business Media.

Blum, A., & Mitchell, T. (1998, July). Combining labeled and unlabeled data with co-training. In Proceedings of the eleventh annual conference on Computational learning theory, pp. 92-100. ACM.

Bollen, N. P., & Whaley, R. E. (2004). Does net buying pressure affect the shape of implied volatility functions?. The Journal of Finance, 59(2), pp. 711-753.

Bormann, S. K. (2013). Sentiment indices on financial markets: What do they measure? (No. 2013-58). Economics Discussion Papers.

Bosch-Domènech, A., & Silvestre, J. (2010). Averting risk in the face of large losses: Bernoulli vs. Tversky and Kahneman. Economics Letters, 107(2), pp. 180-182.

Bram, J., & Ludvigson, S. C. (1997). Does consumer confidence forecast household expenditure? A sentiment index horse race. Federal Reserve Bank of New York Economic Policy Review 4, (2), pp. 59-78.

Breiman, L. (1996). BAGGING predictors. Machine learning, 24(2), pp. 123-140.

Breiman, L. (2001). *RANDOM FOREST*s. Machine learning, 45(1), pp. 5-32.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and regression *TREE*s. CRC press.

Brenning, A. (2009). Benchmarking classifiers to optimally integrate terrain analysis and multispectral remote sensing in automatic rock glacier detection. Remote Sensing of Environment, 113(1), pp. 239-247.

Brown, G. W., & Cliff, M. T. (2005). Investor sentiment and asset valuation. The Journal of Business, 78(2), pp. 405-440.

Brzezicka, J., & Wiśniewski, R. (2013). Calendar effects on the real estate market. Real Estate Management and Valuation, 21(2), pp. 13-21.

Byrne, P., Jackson, C., & Lee, S. (2013). Bias or rationality? The case of UK commercial real estate investment. Journal of European real estate research, 6(1), pp. 6-33.

Camerer, C., Issacharoff, S., Loewenstein, G., O'donoghue, T., & Rabin, M. (2003). Regulation for Conservatives: Behavioral Economics and the Case for" Asymmetric Paternalism". University of Pennsylvania law review, 151(3), pp. 1211-1254.

Carroll, C. D., Fuhrer, J. C., & Wilcox, D. W. (1994). Does consumer sentiment forecast household spending? If so, why?. The American Economic Review, 84(5), pp. 1397-1408.

Case, K. E., Shiller, R. J., & Thompson, A. (2012). What have they been thinking? Homebuyer behaviour in hot and cold markets (No. w18400). National Bureau of Economic Research.

Case, K.E., Shiller, R.J., (1989). The efficiency of the market for single-family homes. The American Economic Review 79 (1), pp. 125–137.

Chen, C. C., & Tseng, Y. D. (2011). Quality evaluation of product reviews using an information quality framework. Decision Support Systems, 50(4), pp. 755-768.

Chen, T., Xu, R., He, Y., Xia, Y., & Wang, X. (2016). Learning user and product distributed representations using a sequence model for sentiment analysis. IEEE Computational Intelligence Magazine, 11(3), pp. 34-44.

Chervachidze, S., & Wheaton, W. (2013). What determined the Great Cap Rate Compression of 2000–2007, and the dramatic reversal during the 2008–2009 Financial Crisis?. The Journal of Real Estate Finance and Economics, 46(2), pp. 208-231.

Chervachidze, S., Costello, J., & Wheaton, W. C. (2009). The secular and cyclic determinants of capitalization rates: the role of property fundamentals, macroeconomic factors, and "structural changes". The Journal of Portfolio Management, 35(5), pp. 50-69.

Chiang, K. C., & Lee, M. L. (2010). The role of correlated trading in setting REIT prices. The Journal of Real Estate Finance and Economics, 41(3), pp. 320-338.

Chichernea, D., Miller, N., Fisher, J., Sklarz, M., & White, B. (2008). A cross-sectional analysis of cap rates by msa. Journal of Real Estate Research, 30(3), pp. 249-292.

Choi, H., & Varian, H. (2009). Predicting initial claims for unemployment benefits. Google Inc, pp. 1-5.

Choi, H., & Varian, H. (2012). Predicting the present with Google Trends. Economic Record, 88(s1), pp. 2-9.

Chomsky, N. (2014). Aspects of the Theory of Syntax (Vol. 11). MIT press.

Chowdhury, G. G. (2003). Natural language processing. Annual review of information science and technology, 37(1), pp. 51-89.

Clayton, J., Ling, D. C., & Naranjo, A. (2009). Commercial real estate valuation: fundamentals versus investor sentiment. The Journal of Real Estate Finance and Economics, 38(1), pp. 5-37.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), pp. 273-297.

Croce, R. M., & Haurin, D. R. (2009). Predicting turning points in the housing market. Journal of Housing Economics, 18(4), pp. 281-293.

Da, Z., Engelberg, J., Gao, P. (2011). In search of attention. The Journal of Finance, 66 (5), pp. 1461-1499.

Das, P. K., Freybote, J., & Marcato, G. (2015b). An investigation into sentiment-induced institutional trading behaviour and asset pricing in the REIT market. The Journal of Real Estate Finance and Economics, 51(2), pp. 160-189.

Das, P., Ziobrowski, A., Coulson, N. E., (2015b). Online information search, market fundamentals and apartment real estate. The Journal of Real Estate Finance and Economics, 51 (4), pp. 480-502.

Dave, K., Lawrence, S., & Pennock, D. M. (2003, May). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In Proceedings of the 12th international conference on World Wide Web (pp. 519-528). ACM.

David, D. E., Lynne, A. M., Han, J., & Foley, S. L. (2010). Evaluation of virulence factor profiling in the characterization of veterinary Escherichia coli isolates. Applied and environmental microbiology, 76(22), pp. 7509-7513.

De Bondt, W. F., Shefrin, H., Muradoglu, G. Y., & Staikouras, S. K. (2008). Behavioural finance: Quo vadis. Journal of Applied Finance, 19(2), pp. 7-21.

De Long, J. B., Shleifer, A., Summers, L. H., & Waldmann, R. J. (1990). Noise trader risk in financial markets. Journal of Political Economy, 98(4), pp. 703-738.

De Neve, J. E., & Fowler, J. H. (2014). Credit card borrowing and the monoamine oxidase A (MAOA) gene. Journal of Economic Behaviour & Organization, 107, pp. 428-439.

DeCoster, G. P., & Strange, W. C. (2012). Developers, herding, and overbuilding. The Journal of Real Estate Finance and Economics, 44(1-2), pp. 7-35.

Devaney, S., Livingstone, N., McAllister, P. and Nanda, A. (2016). Unravelling Liquidity In International Commercial Real Estate Markets. Investment Property Forum (IPF).

Diaz, J. (1997). An investigation into the impact of previous expert value estimates on appraisal judgment. Journal of Real Estate Research, 13(1), pp. 57-66.

Diaz, J., & Hansz, A. (2007). Understanding the behavioural paradigm in property research. Pacific Rim Property Research Journal, 13(1), 16-34.

Diebold, F. & Mariano, R. (1995). “Comparing Predictive Accuracy”, Journal of Business and Economic Statistics, 13(3), pp. 253-63.

Dietzel, M.A., Braun, N., Schäfers, W., 2014. Sentiment-based commercial real estate forecasting with Google Search volume data. Journal of Property Investment & Finance 36 (6), pp. 540–569.

DiPasquale, D., Wheaton, W.C., 1992. The cost of capital, tax reform, and the future of the rental housing market. Journal of Urban Economics 31 (3), pp. 337–359.

Dominitz, J., & Manski, C. F. (2004). How should we measure consumer confidence?. Journal of Economic Perspectives, 18(2), pp. 51-66.

Doran, J. S., Peterson, D. R., & Price, S. M. (2012). Earnings conference call content and stock price: the case of REITs. The Journal of Real Estate Finance and Economics, 45(2), pp. 402-434.

Dua, P., 2008, Analysis of consumers' perceptions of buying conditions for houses. The Journal of Real Estate Finance and Economics, 37 (4), pp. 335-350.

Duca, J. V., & Ling, D. C. (2015). The other (commercial) real estate boom and bust: the effects of risk premia and regulatory capital arbitrage.

Duric, A., & Song, F. (2012). Feature selection for sentiment analysis based on content and syntax models. Decision Support Systems, 53(4), pp. 704-711.

Easaw, J. Z., & Heravi, S. M. (2004). Evaluating consumer sentiments as predictors of UK household consumption behaviour: Are they accurate and useful?. International Journal of Forecasting, 20 (4), pp. 671-681.

Elton, E. J., Gruber, M. J., & Busse, J. A. (1998). Do investors care about sentiment?. The Journal of Business, 71 (4), pp. 477-500.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. The Journal of Finance, 25 (2), pp. 383-417.

Fama, E. F. (1998). Market efficiency, long-term returns, and behavioral finance1. Journal of financial economics, 49 (3), pp. 283-306.

Fan, C. S., & Wong, P. (1998). Does consumer sentiment forecast household spending?: The Hong Kong case. Economics Letters, 58 (1), pp. 77-84.

Fan, T. K., & Chang, C. H. (2011). Blogger-centric contextual advertising. Expert Systems with Applications, 38 (3), pp. 1777-1788.

Fang, J., & Chen, B. (2013). U.S. Patent No. 8, 352, 405. Washington, DC: U.S. Patent and Trademark Office.

Fawcett, T., & Provost, F. (1999, August). Activity monitoring: Noticing interesting changes in behaviour. In Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 53-62). ACM.

Feinerer, I., & Hornik, K. (2008). tm: Text Mining Package. R - package version 0.3.

Feinerer, K. Hornik; Meyer, D. (2008) Text mining infrastructure in R, Journal of Statistical Software, Volume 25, Issue 5: 1{54, March 2008. ISSN 1548-7660. URL http://www.jstatsoft.org/v25/i05

Fernández-Gavilanes, M., Álvarez-López, T., Juncal-Martínez, J., Costa-Montenegro, E., & González-Castaño, F. J. (2016). Unsupervised method for sentiment analysis in online texts. Expert Systems with Applications, 58, pp. 57-75.

Fersini, E., Messina, E., & Pozzi, F. A. (2016). Expressive signals in social media languages to improve polarity detection. Information Processing & Management, 52(1), pp. 20-35.

Finn, Å. N. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs, DTU Informatics, Technical University of Denmark, Lyngby, Denmark, arXiv preprint arXiv:1103.290.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. Psychological Bulletin, 76(5), pp. 378.

Foote, C. L., Gerardi, K. S., & Willen, P. S. (2012). Why did so many people make so many ex-post bad decisions? The causes of the foreclosure crisis (No. w18082). National Bureau of Economic Research.

French, N. (2001). Decision theory and real estate investment: an analysis of the decision-making processes of real estate investment fund managers. Managerial and decision economics, 22(7), pp. 399-410.

Freybote, J. (2016). Real estate sentiment as information for REIT bond pricing. Journal of Property Research, 33(1), pp. 18-36.

Freybote, J., & Seagraves, P. A. (2017) Heterogeneous investor sentiment and institutional real estate investments. Real Estate Economics., 45 (1), pp. 154-176.

Friedman, J. (1996). Another approach to polychotomous classification (Vol. 56). Technical report, Department of Statistics, Stanford University.

Friedman, J., Hastie, T., & Tibshirani, R. (2009). Lasso and elastic-net regularized generalized linear models. In: R - package.

Friedman, M., & Savage, L. J. (1948). The utility analysis of choices involving risk. Journal of Political Economy, 56(4), pp. 279-304.

Froot, K. A., Bhargava, R., Cuipa, E. S., & Arabadjis, J. S. (2014). Multi-Asset Sentiment and Institutional Investor Behaviour: A Cross-Asset Perspective. The Journal of Portfolio Management, 40(4), pp. 144-156.

Frugier, A. (2016). Returns, volatility and investor sentiment: Evidence from European stock markets. Research in International Business and Finance, 38, pp. 45-55.

Fung, G. P. C., Yu, J. X., & Lam, W. (2002, May). News sensitive stock trend prediction. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 481-493). Springer Berlin Heidelberg.

Fung, G. P. C., Yu, J. X., & Lu, H. (2005). The Predicting Power of Textual Information on Financial Markets. IEEE Intelligent Informatics Bulletin, 5(1), pp. 1-10.

Gabrilovich, E., & Markovitch, S. (2009). Wikipedia-based semantic interpretation for natural language processing. Journal of Artificial Intelligence Research, 34, pp. 443-498.

Gallimore, P. (1996). Confirmation bias in the valuation process: a test for corroborating evidence. Journal of Property Research, 13(4), 261-273.

Gallimore, P., & Wolverton, M. (1997). Price-knowledge-induced bias: a cross-cultural comparison. Journal of Property Valuation and Investment, 15(3), 261-273.

Garcia, M. J. R. (2013). Financial education and behavioural finance: new insights into the role of information in financial decisions. Journal of Economic Surveys, 27(2), pp. 297-315.

Giacomini, E. (2011). The Role of Investor Sentiment in the Real Estate Market (Doctoral dissertation, Dissertation, Università Politecnica Delle Marche, Ancona, Italy).

Gidofalvi, G., & Elkan, C. (2001). Using news articles to predict stock price movements. Department of Computer Science and Engineering, University of California, San Diego.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. Nature, 457(7232), pp. 1012-1014.

Godbole, N., Srinivasaiah, M., & Skiena, S. (2007). Large-Scale Sentiment Analysis for News and Blogs. ICWSM, 7(21), pp. 219-222.

Goodman, J. (1994). Using attitude data to forecast housing activity. Journal of Real Estate Research, 9(4), pp. 445-453.

Graham, E., Hall, W., & Schuhmann, P. (2007). Hurricanes, catastrophic risk, and real estate market recovery. Journal of Real Estate Portfolio Management, 13(3), pp. 179-190.

Hall, R. E. (1978). Stochastic implications of the life cycle-permanent income hypothesis: theory and evidence. Journal of political economy, 86(6), pp. 971-987.

Hardin, W. (1999). Behavioural research into heuristics and bias as an academic pursuit: Lessons from other disciplines and implications for real estate. Journal of Property Investment & Finance, 17(4), pp. 333-352.

Hastie, T., & Qian, J. (2014). *GLMENT* Vignette.

He, Y. (2012). Incorporating sentiment prior knowledge for weakly supervised sentiment analysis. ACM Transactions on Asian Language Information Processing (TALIP), 11(2), pp. 4.

He, Y., & Zhou, D. (2011). Self-training from labelled features for sentiment analysis. Information Processing & Management, 47 (4), pp. 606-616.

Heinig, S., & Nanda, A. (2018). Measuring sentiment in real estate–a comparison study. Journal of Property Investment & Finance, 36 (3), pp. 248-258.

Hendershott, P. H., & MacGregor, B.D. (2005a). Investor rationality: evidence from UK property capitalization rates. Real Estate Economics, 33(2), pp. 299-322.

Hendershott, P. H., & MacGregor, B.D. (2005b). Investor rationality: An analysis of NCREIF commercial property data. Journal of Real Estate Research 27 (4), pp. 445–475.

Hengelbrock, J., Theissen, E., & Westheide, C. (2013). Market response to investor sentiment. Journal of Business Finance & Accounting, 40(7-8), pp. 901-917.

Hirshleifer, D. (2001). Investor psychology and asset pricing. The Journal of Finance, 56(4), pp. 1533-1597.

Hirshleifer, D., & Shumway, T. (2003). Good day sunshine: Stock returns and the weather. The Journal of Finance, 58(3), pp. 1009-1032.

Hodgson, G. M. (1998). The approach of institutional economics. Journal of economic literature, 36(1), pp. 166-192.

Hohenstatt, R., Kaesbauer, M., 2014. GECO's Weather Forecast for the U.K. Housing Market: To What Extent Can We Rely on Google Econometrics?. Journal of Real Estate Research, 36 (2), pp. 253-281.

Hosseini, H. (2011). George Katona: A founding father of old behavioural economics. The Journal of Socio-Economics, 40(6), pp. 977-984.

Howrey, E. P. (2001). The predictive power of the index of consumer sentiment. Brookings papers on economic activity, 2001(1), pp. 175-207.

Hsu, C. W., & Lin, C. J. (2002). A comparison of methods for multiclass support vector machines. IEEE Transactions on Neural Networks, 13(2), pp. 415-425.

Hu, M., & Liu, B. (2004, August). Mining and summarizing customer reviews. In Proceedings of the tenth ACMSIGKDD international conference on Knowledge discovery and data mining (pp. 168-177). ACM.

Hui, E. C. M., Wright, J. A., & Yam, S. C. P. (2014). Calendar effects and real estate securities. The Journal of Real Estate Finance and Economics, 49(1), pp. 91-115.

Hui, E. C. M., Zheng, X., & Wang, H. (2013). Investor sentiment and risk appetite of real estate security market. Applied Economics, 45(19), pp. 2801-2807.

Hung, P. H. (2016). Investor sentiment, order submission, and investment performance on the Taiwan Stock Exchange. Pacific-Basin Finance Journal, 39, pp. 124-140.

Hutchison, N., Fraser, P., Adair, A., & Srivatsa, R. (2012). Regime shifts in ex post-UK commercial property risk premiums. Journal of Property Research, 29(3), pp. 247-269.

Iba, W., & Langley, P. (1992). Induction of one-level decision *TREE*s. In Proceedings of the ninth international conference on machine learning (pp. 233-240).

Irresberger, F., Mühlnickel, J., & Weiß, G. N. (2015). Explaining bank stock performance with crisis sentiment. Journal of Banking & Finance, 59, pp. 311-329.

Jin, C., Soydemir, G., & Tidwell, A. (2014). The US housing market and the pricing of risk: Fundamental analysis and market sentiment. Journal of Real Estate Research., 36 (2), pp. 187-219

Jockers, M. (2016). Package '*SYUZHET*'.

Joel-Carbonell, A., & Rottke, N. B. (2009). Efficient markets versus behavioural anomalies: the case of REITs. Journal of Property Investment & Finance, 27(4), pp. 413-424.

Joseph, K., Wintoki, M. B., & Zhang, Z. (2011). Forecasting abnormal stock returns and trading volume using investor sentiment: Evidence from online search. International Journal of Forecasting, 27(4), pp. 1116-1127.

Jurka, T. P., Collingwood L., Boydstun, A. E., Grossman E. and van Atteveldt W. (2012). RTextTools: Automatic Text Classification via Supervised Learning. R - package version 1.3.9. http://CRAN.R-project.org/package=RTextTools.

Jurka, T. P., Collingwood, L., Boydstun, A. E., Grossman, E., & van Atteveldt, W. (2013). RTextTools: A supervised learning package for text classification. The R Journal, 5(1), pp. 6-12.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. Cognitive psychology, 3(3), pp. 430-454.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. Econometrica: Journal of the econometric society, pp. 263-291.

Kaplanski, G., & Levy, H. (2012). Real estate prices: An international study of seasonality's sentiment effect. Journal of Empirical Finance, 19(1), pp. 123-146.

Katona, G. (1953). Rational behaviour and economic behaviour. Psychological Review, 60(5), pp. 307.

Katona, G. (1968). Consumer behaviour: Theory and findings on expectations and aspirations. The American Economic Review, 58(2), pp. 19-30.

Katz, E. (1957). The two-step flow of communication: An up-to-date report on an hypothesis. Public opinion quarterly, 21(1), pp. 61-78.

Katz, J. J., & Fodor, J. A. (1963). The structure of a semantic theory. language, 39(2), pp. 170-210.

Kauer, A. U., & Moreira, V. P. (2016). Using information retrieval for sentiment polarity prediction. Expert Systems with Applications, 61, pp. 282-289.

Keller, F., & Lapata, M. (2003). Using the web to obtain frequencies for unseen bigrams. Computational linguistics, 29(3), pp. 459-484.

Kiritchenko, S., & Mohammad, S. M. (2016). The effect of negators, modals, and degree adverbs on sentiment composition. In Proceedings of NAACL-HLT, pp. 43-52.

Kishore, R. (2004). Theory of behavioural finance and its application to property market: a change in paradigm. Australian Property Journal, 38(2), 105.

Kumar, A., & Lee, C. (2006). Retail investor sentiment and return comovements. The Journal of Finance, 61(5), pp. 2451-2486.

Kumar, M. A., & Gopal, M. (2009). Least squares twin support vector machines for pattern classification. Expert Systems with Applications, 36(4), pp. 7535-7543.

Kurov, A. (2010). Investor sentiment and the stock market's reaction to monetary policy. Journal of Banking & Finance 34 (1), pp. 139–149.

Labidi, C., & Yaakoubi, S. (2016). Investor sentiment and aggregate volatility pricing. The Quarterly Review of Economics and Finance, 61, pp. 53-63.

Läuter, J. (1992). Stabile multivariate Verfahren: Diskriminanzanalyse, Regressionsanalyse, Faktoranalyse (Vol. 81). VCH.

Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D., & Allan, J. (2000, November). Language models for financial news recommendation. In Proceedings of the ninth international conference on Information and knowledge management (pp. 389-396). ACM.

Lee, C., Shleifer, A., & Thaler, R. H. (1991). Investor sentiment and the closed-end fund puzzle. The Journal of Finance, 46(1), pp. 75-109.

Lee, K., & Timmons, R. (2007). Predicting the stock market with news articles. CS224n Final Report.

Lees, R. B., & Chomsky, N. (1957). Syntactic structures. Language, 33(3 Part 1), pp. 375-408.

Levy, D. (1997) The Impact of the Examination of a Property on the Perception of Value and Desirability of a Following Property, paper presented at RICS Cutting Edge Property Research Conference, Dublin.

Levy, D., & Schuck, E. (1999). The influence of clients on valuations. Journal of Property Investment & Finance, 17(4), 380-400.

Liang, W. L. (2016). Sensitivity to investor sentiment and stock performance of open market share repurchases. Journal of Banking & Finance, 71, pp. 75-94.

Liaw, A., & Wiener, M. (2002). Classification and regression by random*RANDOM FOREST*. R News, 2(3), pp. 18-22.

Liddy, E. D. (2001). Natural language processing.

Lin, C. Y., Rahman, H., & Yung, K. (2009). Investor sentiment and REIT returns. The journal of real estate finance and economics, 39(4), p. 450.

Lin, C., He, Y., Everson, R., & Ruger, S. (2012). Weakly supervised joint sentiment-topic detection from text. IEEE Transactions on Knowledge and Data Engineering, 24 (6), pp. 1134-1145.

Lin, Y. H. C., Gao, W., & Wong, K. F. (2012). Tracking sentiment and topic dynamics from social media.

Ling, D. C., Naranjo, A., & Scheick, B. (2014). Investor sentiment, limits to arbitrage and private market returns. Real Estate Economics, 42(3), pp. 531-577.

Lintner, J. (1965). The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. The review of economics and statistics, pp. 13-37.

Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), pp. 1-167.

Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), pp. 1-167.

Liu, B., Hu, M., & Cheng, J. (2005, May). Opinion observer: analyzing and comparing opinions on the web. In Proceedings of the 14th international conference on World Wide Web (pp. 342-351). ACM.

Liu, B., Hu, M., & Cheng, J. (2005, May). Opinion observer: analyzing and comparing opinions on the web. In Proceedings of the 14th international conference on World Wide Web (pp. 342-351). ACM.

Liu, B., Li, X., Lee, W. S., & Yu, P. S. (2004, July). Text classification by labeling words. In AAAI (Vol. 4, pp. 425-430).

Loughlin, C., & Harnisch, E. (2014). The viability of StockTwits and Google Trends to predict the stock market. Retrieved from stocktwits: http://stocktwits. com/research/Viability-of-StockTwits-and-Google-Trends-Loughlin_Harnisch. pdf.

Loughran, T., & McDonald, B. (2014). Measuring readability in financial disclosures. The Journal of Finance, 69(4), pp. 1643-1671.

MacCowan, R. J., & Orr, A. M. (2008). A behavioural study of the decision processes underpinning disposals by property fund managers. Journal of Property Investment & Finance, 26(4), pp. 342-361.

Maks, I., & Vossen, P. (2012). A lexicon model for deep sentiment analysis and opinion mining applications. Decision Support Systems, 53(4), pp. 680-688.

Malgarini, M., & Margani, P. (2007). Psychology, consumer sentiment and household expenditures. Applied Economics, 39(13), pp. 1719-1729.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014, June). The Stanford core nlp natural language processing toolkit. In ACL (System Demonstrations), pp. 55-60.

Marcato, G., & Nanda, A. (2016). Information content and forecasting ability of sentiment indicators: case of real estate market. Journal of Real Estate Research, 38(2), pp. 165-203.

Markowitz, H. (1952a). Portfolio selection. The journal of finance, 7(1), pp. 77-91.

Markowitz, H. (1952b). The utility of wealth. Journal of Political Economy, 60(2), pp. 151-158.

Mayhew, S., & Stivers, C. (2003). Stock return dynamics, option volume, and the information content of implied volatility. Journal of Futures Markets, 23(7), pp. 615-646.

Maynard, D., & Bontcheva, K. (2016, May). Challenges of Evaluating Sentiment Analysis Tools on Social Media. In Proceedings of LREC 2016. LREC.

Maynard, D., & Funk, A. (2011, May). Automatic detection of political opinions in tweets. In Extended Semantic Web Conference (pp. 88-99). Springer Berlin Heidelberg.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. Biochemia Medica, 22(3), pp. 276-282.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5 (4), pp. 1093-1113.

Merton, R. K. (1948). The self-fulfilling prophecy. The Antioch Review, 8(2), pp. 193-210.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2014). e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. R - package version 1.6-3.

Mian, G. M., & Sankaraguruswamy, S. (2012). Investor sentiment and stock market response to earnings news. The Accounting Review, 87(4), pp. 1357-1384.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychological Review, 63(2), p. 81.

Mohammad, S. M., & Turney, P. D. (2010, June). Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text (pp. 26-34). Association for Computational Linguistics.

Montoyo, A., MartíNez-Barco, P., & Balahur, A. (2012). Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments.

Mossin, J. (1966). Equilibrium in a capital asset market. Econometrica: Journal of the econometric society, pp. 768-783.

Mudinas, A., Zhang, D., & Levene, M. (2012, August). Combining lexicon and learning based approaches for concept-level sentiment analysis. In Proceedings of the first international workshop on issues of sentiment discovery and opinion mining (p.5). ACM.

Nanda, A. (2007). Examining the NAHB/Wells Fargo Housing Market Index (HMI). Housing Economics.

Nasukawa, T., & Yi, J. (2003, October). Sentiment analysis: Capturing favorability using natural language processing. In Proceedings of the 2nd international conference on Knowledge Capture (pp. 70-77). ACM.

Nguyen, T. H., Shirai, K., & Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. Expert Systems with Applications, 42 (24), pp. 9603-9611.

Nielsen, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. arXiv preprint arXiv:1103.290.

Nigam, K., Lafferty, J., McCallum, A. (1999). Using Maximum Entropy for Text Classification, School of Computer Science, Carnegie Mellon University, Pittsburgh.

Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. (2000). Machine Learning; Chapter: Text classification from labeled and unlabeled documents using EM. Machine learning, Kluwer Academic Publishers, Boston. Manufactured in The Netherlands, 39 (2-3), pp. 103-134.

Nisbett, R. E., & Wilson, T. D. (1977). The halo effect: Evidence for unconscious alteration of judgments. Journal of personality and social psychology, 35(4), pp. 250-256.

Northcraft, G. B., & Neale, M. A. (1987). Experts, amateurs, and real estate: An anchoring-and-adjustment perspective on property pricing decisions. Organizational behaviour and human decision processes, 39(1), pp. 84-97.

O'Hare, N., Davy, M., Bermingham, A., Ferguson, P., Sheridan, P., Gurrin, C., & Smeaton, A. F. (2009, November). Topic-dependent sentiment analysis of financial blogs. In Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion (pp. 9-16). ACM.

O'Keefe, T., Curran, J. R., Ashwell, P., & Koprinska, I. (2013). An annotated corpus of quoted opinions in news articles. In ACL (2) (pp. 516-520).

Okugami, C. (2013) Google trend data downloading API for R, Package name GOOGLETREND, Published on Github.

Palmer, D. D. (2010). Chapter 2 – Text Preprocessing in The Handbook of Natural Language Processing. Second Edition, edited by Indurkhya, Nitin; Damerau, Fred J., CRC Press – Taylor and Francis Group.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval, 2(1–2), pp. 1-135.

Pang, B., Lee, L., & Vaithyanathan, S. (2002, July). Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing, Volume 10 (pp. 79-86). Association for Computational Linguistics.

Park, H., & Sohn, W. (2013). Behavioural finance: A survey of the literature and recent development. Seoul Journal of Business, 19(1), pp. 3-42.

Peters, A., Hothorn, T., Ripley, B. D., Therneau, T., & Atkinson, B. ipred: Improved Predictors, 2013. R - package version 0.9-4.

Porter, M. F. (1980). An algorithm for suffix stripping. Program, 14(3), pp. 130-137.

Posner, R. A. (2012). Behavioural Finance before Kahneman. Loy. U. Chi. LJ, 44, 1341.

Preis, T., Moat, H. S., & Stanley, H. E. (2013). Quantifying trading behavior in financial markets using Google Trends. Scientific reports, 3, 1684.

Preis, T., Moat, H. S., Stanley, H. E., & Bishop, S. R. (2012). Quantifying the advantage of looking forward. Scientific reports, 2, 350.

Preis, T., Reith, D., & Stanley, H. E. (2010). Complex dynamics of our economic life on different scales: insights from search engine query data. Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, 368(1933), pp. 5707-5719.

Pressman, S. (2006). Kahneman, Tversky, and institutional economics. Journal of Economic Issues, 40(2), pp. 501-506.

Rabin, M. (1998). Psychology and economics. Journal of economic literature, 36(1), pp. 11-46.

Rajakumari, S. B. (2014). Data Quality Mining in Electronic News Paper. Indian Journal of Science and Technology, 7(S5), pp. 47-50.

Rao, C. R. (1948). The utilization of multiple measurements in problems of biological classification. Journal of the Royal Statistical Society. Series B (Methodological), 10(2), pp. 159-203.

Ratcliff, R. U. (1972). Valuation for real estate decisions. Democrat Press.

Ricciardi, V., & Simon, H. K. What is Behavioral Finance?. Business, Education & Technology Journal, 2 (2), pp. 1-9.

Ripley, B. (2007). *TREE*: classification and regression *TREE*s. [R - package version 1.0-26].

Ripley, B., & Venables, W. (2011). *NNET*: Feed-forward neural networks and multinomial log-linear models. R - package version, 7(5).

Ross, S. A. (1976). The arbitrage theory of capital asset pricing. Journal of economic theory, 13(3), pp. 341-360.

Sadique, S., In, F. H., & Veeraraghavan, M. (2008). The impact of spin and tone on stock returns and volatility: Evidence from firm-issued earnings announcements and the related press coverage.

Saif, H., He, Y., Fernandez, M., & Alani, H. (2016). Contextual semantics for sentiment analysis of Twitter. Information Processing & Management, 52(1), pp. 5-19.

Savoy, James; Gaussier, Eric (2010). Chapter 19 – Information Retrieval in The Handbook of Natural Language Processing. Second Edition. edited by Indurkhya, N., Damerau, F. CRC Press – Taylor and Francis Group.

Schapire, R. E., & Freund, Y. (2012). *BOOSTING*: Foundations and algorithms. MIT press.

Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The AZFin text system. ACM Transactions on Information Systems (TOIS), 27(2), 12.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. The journal of finance, 19(3), pp. 425-442.

Shefrin, H. (2000). Beyond Greed and Fear Harvard Business School Press.

Sheu, H. J., & Wei, Y. C. (2011). Effective options trading strategies based on volatility forecasting recruiting investor sentiment. Expert Systems with Applications, 38(1), pp. 585-596.

Shiller, R. J. (2003). From efficient markets theory to behavioural finance. The Journal of Economic Perspectives, 17(1), pp. 83-104.

Shiller, R. J., Fischer, S., & Friedman, B. M. (1984). Stock prices and social dynamics. Brookings papers on economic activity, 1984(2), pp. 457-510.

Shilling, J. D., & Sing, T. F. (2007, March). Do institutional real estate investors have rational expectations. In Asian Real Estate Society (AsRES) Annual Conference Paper.

Simon, H. A. (1957). Models of man; social and rational.

Sivitanides, P., Southard, J., Torto, R. G., & Wheaton, W. C. (2001). The determinants of appraisal-based capitalization rates. Real Estate Finance, 18(2), pp. 27-38.

Sivitanidou, R., & Sivitanides, P. (1999). Office capitalization rates: Real estate and capital market influences. The Journal of Real Estate Finance and Economics, 18(3), pp. 297-322.

Smales, L. A. (2016). News sentiment and bank credit risk. Journal of Empirical Finance, 38, pp. 37-61.

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013, October). Recursive deep models for semantic compositionality over a sentiment *Tree*bank. In Proceedings of the conference on empirical methods in natural language processing (EMNLP) (Vol. 1631, pp. 1631 - 1642).

Soo, C. K. (2015). Quantifying animal spirits: news media and sentiment in the housing market.

Soroka, S., & McAdams, S. (2015). News, politics, and negativity. Political Communication, 32(1), pp. 1-22.

Sprenger, T. O., Tumasjan, A., Sandner, P. G., & Welpe, I. M. (2014). Tweets and trades: The information content of stock microblogs. European Financial Management, 20(5), pp. 926-957.

Statman, M. (1995, December). Behavioural finance versus standard finance. In AIMR Conference Proceedings (Vol. 1995, No. 7, pp. 14-22). Association for Investment Management and Research.

Steinberger, J., Ebrahim, M., Ehrmann, M., Hurriyetoglu, A., Kabadjov, M., Lenkova, P., & Zavarella, V. (2012). Creating sentiment dictionaries via triangulation. Decision Support Systems, 53(4), pp. 689-694.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. The Journal of Finance, 62(3), pp. 1139-1168.

Tetlock, P. C. (2011). All the news that's fit to reprint: Do investors react to stale information?. Review of Financial Studies, 24(5), pp. 1481-1512.

Tetlock, P. C., SAAR-TSECHANSKY, M. A. Y. T. A. L., & Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. The Journal of Finance, 63(3), pp. 1437-1467.

Tetlock, P. C., Saar-Tsechansky, M., & Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. The Journal of Finance, 63(3), pp. 1437-1467.

Thaler, R. H. (2010). The end of behavioural finance.

Treynor, J. L., (1962). Toward a Theory of Market Value of Risky Assets, Unpublished manuscript, final version was published in 1999, in Asset Pricing and Portfolio Performance: Models, Strategy and Performance Metrics. Robert A. Korajczyk (editor) London: Risk Books, pp. 15–22

Tsai, F. T., Lu, H. M., & Hung, M. W. (2016). The impact of news articles and corporate disclosure on credit risk valuation. Journal of Banking & Finance, 68, pp. 100-116.

Tsolacos, S. (2006). An assessment of property performance forecasts: consensus versus econometric. Journal of Property Investment & Finance, 24(5), pp. 386-399.

Tsolacos, S. (2012). The role of sentiment indicators for real estate market forecasting. Journal of European Real Estate Research, 5(2), pp. 109-120.

Tsolacos, S., Brooks, C., & Nneji, O. (2014). On the predictive content of leading indicators: the case of US real estate markets. Journal of Real Estate Research, 36(4), 541-573.

Turing, A. M. (1950). Computing machinery and intelligence. Mind, 59(236), pp. 433-460.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. Psychological Bulletin, 76(2), p. 105.

Vohra, S. M., & Teraiya, J. (2013). A comparative study of sentiment analysis techniques. Journal JIKRCE, 2 (2), pp. 313-317.

Vosen, S., & Schmidt, T. (2011). Forecasting private consumption: survey-based indicators vs. Google trends. Journal of Forecasting, 30(6), pp. 565-578.

Walker, C. B. (2014a) Media and Opinion Leaders in the Housing Market, Queen's University Belfast, Working Paper FIN 14-8

Walker, C. B. (2014b). Housing booms and media coverage. Applied Economics, 46(32), pp. 3954-3967.

Walker, C. B. (2016). The direction of media influence: Real-estate news and the stock market. Journal of Behavioural and Experimental Finance, 10, pp. 20-31.

Walker, M. A., Anand, P., Abbott, R., *TREE*, J. E. F., Martell, C., & King, J. (2012). That is your evidence?: Classifying stance in online political debate. Decision Support Systems, 53(4), pp. 719-729.

Weber, W., & Devaney, M. (1996). Can consumer sentiment surveys forecast housing starts?. Appraisal Journal, 64, pp. 343-350.

Welling, M. (2005). Fisher linear discriminant analysis. Department of Computer Science, University of Toronto, 3, pp. 1-4.

Wofford, L. E. (1985). Cognitive processes as determinants of real estate investment decisions. Appraisal Journal, 53(July), pp. 388-395.

Wofford, L., Troilo, M., & Dorchester, A. (2011). Point of View: Cognitive Risk and Real Estate Portfolio Management. Journal of Real Estate Portfolio Management, 17(1), pp. 69-73.

Wu, L., & Brynjolfsson, E. (2015). The future of prediction: How Google searches foreshadow housing prices and sales. In Economic analysis of the digital economy (pp. 89-118). University of Chicago Press.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., & Klingner, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144.

Xiao, R. (2010). Chapter 2 – Corpus Creation in The Handbook of Natural Language Processing. Second Edition. edited by Indurkhya, N., Damerau, F. CRC Press – Taylor and Francis Group.

Yang, C. Y., Jhang, L. J., & Chang, C. C. (2016). Do investor sentiment, weather and catastrophe effects improve hedging performance? Evidence from the Taiwan options market. Pacific-Basin Finance Journal, 37, pp. 35-51.

Yan-Yan, Z., *BING*, Q., & Ting, L. (2010). Integrating intra-and inter-document evidences for improving sentence sentiment classification. Acta Automatica Sinica, 36(10), pp. 1417-1425

8 APPENDIX

CHAPTER 3 - SENTIMENT PROXIES

Table 8:1 - Scoring coefficients (macroeconomic sentiment - Kaiser Criterion)

Labels	Component 1	Component 2	Component 3	Component 4	Component 5
Standardized residual of the ESI	0.287	0.205	-0.550	-0.170	0.039
Standardized residual of the ESI (1 lag)	0.296	0.199	-0.549	-0.155	-0.063
Standardized residual of the change of the stockmarket return	0.012	0.009	0.005	-0.166	0.817
Standardized residual of the change of the stockmarket return (1 lag)	0.029	0.026	-0.024	-0.210	-0.556
Standardized residual of the change of consumer confidence	0.144	0.423	0.361	-0.396	0.020
Standardized residual of the change of consumer confidence (1 lag)	0.151	0.422	0.358	-0.390	-0.038
Standardized residual of the credit rating	0.405	-0.247	0.251	0.083	-0.053
Standardized residual of the credit rating (1 lag)	0.397	-0.255	0.257	0.088	-0.050
Standardized residual of the 10-year government bond rate	-0.179	0.440	0.047	0.417	-0.018
Standardized residual of the 10-year government bond rate (1 lag)	-0.177	0.452	0.040	0.394	-0.029
Standardized residual of the BCI	0.446	0.130	0.028	0.331	0.083
Standardized residual of the BCI (1 lag)	0.445	0.13	0.030	0.332	0.027

Note 8.1: The table provides the correlation coefficients for the 6 times 2 residuals and the identified 5 components from the PCA.

Table 8:2 - Correlation between the various residuals and the components (macroeconomic sentiment - Kaiser Criterion)

Labels	Component 1	Component 2	Component 3	Component 4	Component 5
Standardized residual of the ESI	0.522	0.340	-0.710	-0.192	0.039
Standardized residual of the ESI (1 lag)	0.538	0.329	-0.708	-0.174	-0.064
Standardized residual of the change of the stock market return	0.024	0.015	0.007	-0.186	0.825
Standardized residual of the change of the stock market return (1 lag)	0.054	0.044	-0.032	-0.236	-0.561
Standardized residual of the change of consumer confidence	0.263	0.700	0.466	-0.444	0.021
Standardized residual of the change of consumer confidence (1 lag)	0.275	0.698	0.463	-0.437	-0.039
Standardized residual of the credit rating	0.735	-0.409	0.324	0.093	-0.054
Standardized residual of the credit rating (1 lag)	0.721	-0.421	0.333	0.099	-0.051
Standardized residual of the 10-year government bond rate	-0.326	0.728	0.061	0.468	-0.019
Standardized residual of the 10-year government bond rate (1 lag)	-0.321	0.748	0.053	0.442	-0.029
Standardized residual of the BCI	0.811	0.216	0.037	0.372	0.085
Standardized residual of the BCI (1 lag)	0.809	0.215	0.040	0.373	0.027

Note 8.2: The table illustrates the correlation between the various residuals and the five identified components from the PCA. The correlations are used to identify if a lagged or unlagged residual will be used to construct the sentiment measure. The residual with the highest correlation value will be used.

Table 8:3 - Correlation analysis (macroeconomic sentiment - Kaiser Criterion)

Labels	Component 1	Component 2	Component 3	Component 4	Component 5
Temporary sentiment indicators	0.994	0.984	0.992	0.948	0.813

Note 8.3: The table provides the correlation between the temporary sentiment indicator and the 5 identified components, from the Kaiser Criterion.

Table 8:4 - Calculated weight for final sentiment construction (macroeconomic sentiment - Kaiser Criterion)

	Proportion	Weight
Component 1	0.274	0.331
Component 2	0.227	0.274
Component 3	0.139	0.167
Component 4	0.105	0.126
Component 5	0.085	0.102
Total	0.830	1.000

Note 8.4: The table illustrates the final construction of the macroeconomic sentiment measure, following the Kaiser Criterion. Different to the suggested method, the Kaiser Criterium suggest the use of all Components, which have an eigenvalue above one.

Table 8:5 - PCA of the sentiment proxies (macroeconomic sentiment - PCA)

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	1.779	0.207	0.297	0.297
Comp2	1.572	0.655	0.262	0.559
Comp3	0.917	0.039	0.153	0.711
Comp4	0.878	0.435	0.146	0.858
Comp5	0.442	0.030	0.074	0.931
Comp6	0.412	.	0.069	1.000

Note 8.5: The table illustrates the PCA for the macroeconomic sentiment measure. In total six components have been generated, while naturally the first component has the highest eigenvalue and provides the largest share.

Table 8:6 - Scoring coefficients (macroeconomic sentiment - PCA)

Labels	Component 1	Component 2	Component 3	Component 4	Component 5	Component 6
Standardized residual of the ESI	0.563	0.204	-0.344	-0.280	0.563	-0.357
Standardized residual of the change of the stock market return	0.302	0.143	0.894	-0.278	0.091	0.057
Standardized residual of the change of consumer confidence	0.380	-0.013	0.146	0.903	0.128	-0.055
Standardized residual of the credit rating	-0.241	0.655	-0.068	0.103	0.387	0.590
Standardized residual of the 10-year government bond rate	0.250	-0.651	-0.083	-0.104	0.290	0.642
Standardized residual of the BCI	0.572	0.293	-0.221	-0.089	-0.652	0.325

Note 8.6: The table provides all scoring coefficients for the PCA of the macroeconomic sentiment measure.

Table 8:7 - Orthogonalization process (office sentiment II)

Variables	Labels	IPD: total return index (office)
logofr	logofr	130.066***
		[20.470]
Observations		2,519
R-squared		0.416
Adjusted R-squared		0.416
F-statistics		40.37
Degrees of freedom		64
Number of clusters		65

Robust standard errors in brackets; *** p<0.01, ** p<0.05, * p<0.1

Note 8.7: The table displays the orthogonalization process for the office sentiment II measure. Similar to original retail measure only the log of the office rent has been used.

Table 8:8 - PCA of the sentiment proxies (property sentiment I)

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	3.394	2.902	0.849	0.849
Comp2	0.492	0.380	0.123	0.972
Comp3	0.113	0.112	0.028	1.000
Comp4	0.001	.	0.000	1.000

Note 8.8: The table illustrates the PCA for the property sentiment I. In total four components and there Eigenvalues were used for the construction.

Table 8:9 - Scoring coefficients for all components (property sentiment I)

Labels	Component 1	Component 2	Component 3	Component 4
Office sentiment	0.493	0.481	-0.723	-0.021
Office sentiment (1 lag)	0.487	0.536	0.688	0.022
Retail sentiment	0.509	-0.491	-0.000	0.706
Retail sentiment (1 lag)	0.509	-0.488	0.043	-0.706

Note 8.9: The table provides the scoring coefficients for all components from the PCA.

Table 8:10 - Correlation analysis (property sentiment I)

Variable	Labels	Correlation
pc1(e)	First component	1.000
office_sen~t	Office sentiment	0.909
loffice_se~t	Office sentiment (1 lag)	0.897
retail_sen~t	Retail sentiment	0.939
lretail_se~t	Retail sentiment (1 lag)	0.939

Note 8.10: The table provides the correlation between the sentiment proxies and the first component for the construction of the property sentiment I measure.

Table 8:11 - Variable definition for the yield models

Variable name	Variable definition	Source	Expected sign
ofy	Log of the quarterly office yield	Cushman & Wakefield (formerly DTZ)	
rety	Log of the quarterly retail yield	Cushman & Wakefield (formerly DTZ)	
gbondr	10-year national government bond rate	Datastream	+
rprem	The risk premium is calculated as an eight-quarter rolling standard deviation from the national stock market return	Constructed	+
expected_rent_office	Four-quarter moving average of the deviation of the log of real office rent (Hendershott approach)	Constructed based on Cushman and Wakefield (formerly DTZ) rent data	–
expected_rent_retail	Four-quarter moving average of the deviation of the log of real retail rent (Hendershott approach)	Constructed based on Cushman and Wakefield (formerly DTZ) rent data	–

Note 8.11: The table provides the definition and sources of the used variables.

Table 8:12 - Data description

Variable name	Variable labels
ofy	Office yield
rety	Retail yield
ofr	Office rent
retr	Retail rent
Expected_rent_office	Four-quarter moving average of the deviation of the log of real office rent
Expected_rent_retail	Four-quarter moving average of the deviation of the log of real retail rent
gdp	GDP
fc_gdp	Forecasted change of GDP by the EU and IMF
c_gdp	Change of GDP
cpi	Consumer price index
unemp	Unemployment rate
cred	Credit rating
ipdtroff	IPD total return office
ipdtrret	IPD total return retail
stoind	Stock index
gbondr	Government bond
rprem	Risk premium
intr	Interest rate
csp	Consumer spending
indpropc	Industry production percentage change
esi	Economic sentiment index by the European Union
bci	Business cycle index by the European Union
hcpi	Harmonized consumer price index (EU)

Note 8.12: This table reports all the used variables within this panel dataset and the corresponding acronyms.

Table 8:13 - Descriptive statistics (1)

Variable		Mean	Std. dev.	Min	Max	Observations
Office yield	overall	6.151	1.577	3.500	20.000	N = 3014
	between		1.471	3.951	13.066	n = 74
	within		0.740	2.285	13.085	T-bar = 40.729
Retail yield	overall	5.856	1.857	2.500	19.000	N = 2272
	between		1.724	3.531	12.327	n = 58
	within		0.853	2.877	13.377	T-bar = 39.172
Office rent	overall	33.389	21.110	9.000	185.486	N = 3170
	between		20.443	10.138	142.826	n = 77
	within		5.709	-14.678	78.626	T-bar = 41.168
Retail rent	overall	227.629	214.435	14.480	1,666.670	N = 2222
	between		205.435	14.480	993.687	n = 57
	within		63.926	-76.248	923.755	T-bar = 38.982
Expected rent (office)	overall	-0.189	0.636	-3.475	0.875	N = 3380
	between		0.512	-2.670	0.007	n = 77
	within		0.381	-3.022	2.923	T-bar = 43.896
Expected rent (retail)	overall	-0.359	0.981	-4.504	0.744	N = 2508
	between		0.810	-3.428	0.004	n = 57
	within		0.564	-4.226	3.466	T = 44
GDP	overall	307,332.000	230,490.000	3,259.000	685,900.000	N = 3484
	between		231,147.000	3,989.000	644,427.000	n = 80
	within		25,620.000	223,647.000	395,065.000	T-bar = 43.55
Forecasted change of GDP	overall	0.005	0.006	-0.072	0.109	N = 3520
	between		0.002	0.003	0.013	n = 80
	within		0.006	-0.073	0.102	T = 44
Change of GDP	overall	0.004	0.042	-0.273	0.246	N = 3480
	between		0.005	-0.011	0.023	n = 80
	within		0.042	-0.291	0.261	T-bar = 43.5
Consumer price index	overall	88.827	128.247	-6.090	1,209.600	N = 3520
	between		127.537	1.539	1,022.309	n = 80
	within		19.506	-142.915	276.118	T = 44
Unemployment rate	overall	7.131	3.635	1.100	26.940	N = 3497
	between		3.006	2.027	16.589	n = 80
	within		2.065	-1.528	17.482	T-bar = 43.712
Credit rating	overall	17.853	4.001	0.001	20.000	N = 3494
	between		3.629	4.901	20.000	n = 80
	within		1.818	1.425	22.293	T = 43.675
IPD Total return (office)	overall	438.217	558.043	-2.748	1,985.860	N = 2785
	between		540.433	3.648	1,290.901	n = 68
	within		138.749	50.761	1,133.176	T-bar = 40.955

Note 8.13: The table illustrates the descriptive statistics.

Table 8:14 - Descriptive statistics (2)

Variable		Mean	Std. dev.	Min	Max	Observations
IPD total return (retail)	overall	578.334	755.645	-3.225	2376.150	N = 2780
	between		741.602	7.696	1795.432	n = 68
	within		142.607	63.359	1159.052	T-bar = 40.882
Stock market index	overall	135988.000	227690.000	15.000	680292.000	N = 3334
	between		226091.000	33.000	562018.000	n = 76
	within		35867.000	-30469.000	254263.000	T-bar = 43.868
10-year government bond rate	overall	3.816	1.763	0.310	14.020	N = 3378
	between		1.507	0.537	9.066	n = 79
	within		1.197	-0.105	12.655	T-bar = 42.759
Risk premium	overall	9.004	4.528	2.170	30.447	N = 3202
	between		2.235	6.652	18.454	n = 75
	within		4.048	-2.142	22.232	T-bar = 42.693
National interest rate	overall	2.812	3.086	-0.750	22.000	N = 3520
	between		2.332	0.744	11.016	n = 80
	within		2.037	-6.350	15.835	T = 44
Consumer spending	overall	182994.000	137935.000	1661.000	407413.000	N = 3482
	between		137900.000	2103.000	364750.000	n = 80
	within		18566.000	125798.000	242845.000	T-bar = 43.525
Industry production	overall	0.097	2.531	-18.700	13.300	N = 3505
	between		0.488	-0.552	1.286	n = 80
	within		2.484	-18.571	12.681	T-bar = 43.812
Economic sentiment index	overall	98.858	16.419	-58.200	118.800	N = 3308
	between		12.894	-11.323	104.011	n = 76
	within		10.182	51.980	128.180	T-bar = 43.526
Business climate index	overall	100.116	1.533	85.100	108.633	N = 3412
	between		0.360	98.668	101.197	n = 80
	within		1.490	85.477	108.977	T = 42.65
Harmonized consumer price index (EU)	overall	111.064	12.162	89.827	210.867	N = 3426
	between		6.550	102.812	143.086	n = 78
	within		10.276	57.805	178.845	T-bar = 43.923

Note 8.14: The table represents the descriptive statistics.

Table 8:15 - Google Trends indicator construction

Search words	Total frequency per word	Search words	Total frequency per word	Search words	Total frequency per word
REIT	7	Cushman and Wakefield	2	Royal Bank of Scotland	1
Rent	51	Knight Frank	10	Societe Generale	6
real estate	49	office lease	5	Banco Santander	2
Debt	11	office rent	12	Lloyds Bank	7
Sale	50	office for sale	4	ING	22
Investment	23	office rental	9	UBS	8
Investor	8	commercial office space	1	UniCredit	5
Credit	30	office	41	Credit Suisse	2
Boom	4	office space	8	Rabobank	4
Bust	5	retail	12	Nordea	7
Raise	10	retail space	6	BBVA	6
increase	7	retail rent	2	Commerzbank	7
decrease	3	retail for sale	1	Credit Mutuel	4
shopping centre	18	commercial retail	3	KfW	5
high street	11	retail lease	1	Danske Bank	4
finance	23	retail property	6	Sberbank of Russia	0
mortgage	25	Newmark Grubb Knight Frank	0	CaixaBank	0
loan	16	BNP	10	Handelsbanken	3
commercial real estate	6	BNP real estate	2	Dexia	1
commercial property	15	CoStar	0	KBC	3
commercial property sale	10	Blackstone	2	Nationwide	8
property for sale	26	RE/MAX	0	Bankia	2
lease commercial property	3	Prudential	8	Swedbank	5
commercial lease	9	Voit Real Estate Services	0	La Banque Postale	4
JLL	6	Century 21 Real Estate LLC	0	VTB	2
CBRE	11	HSBC	16	Banco Sabadell	4
Jones Lang LaSalle	12	BNP Paribas	7	Bank of Ireland	0
Colliers	4	Credit Agricole	7	Deka	1
Savills	11	Barclays	15	CB Richard Ellis	2
DTZ	15	Deutsche Bank	9	City name	51

Note 8.15: The table illustrates the overall frequency of the search words for the online search volume index.

Table 8:16 - Google Trends results for each city region

Region	Sum of words	Region	Sum of words
Antwerp	7	Rotterdam	11
Brussels	12	The Hague	9
Liège	5	Utrecht	9
Prague*	27	Oslo*	30
Aarhus	5	Kraków	9
Copenhagen	7	Warsaw	13
Triangle Area	4	Bucharest	23
Helsinki*	25	Moscow	12
Paris	31	Barcelona	14
Lyon	19	Madrid	20
Marseille	19	Gothenburg	5
Berlin (region)	3	Malmö	4
Berlin (city share)	25	Stockholm	7
Düsseldorf	14	Geneva	4
Frankfurt	24	Zürich	8
Hamburg (Region)	3	Istanbul	13
Hamburg (city share)	24	Birmingham	32
Munich	22	Bristol	17
Budapest	8	Leeds	14
Cork	10	London	57
Dublin	22	Manchester	36
Galway	6	Newcastle	6
Limerick	6	Nottingham	8
Milan	18	Sheffield	18
Rome	17	Cardiff	16
Riga	15	Edinburgh	24
Luxembourg City*	31	Glasgow	23
Amsterdam	11		

* National wide search

Note 8.16: This table illustrates the regions within the panel and how many search words out of the 90 have contributed to the regional indicator.

Table 8:17 - Regional fixed effects for the office yield model (1)

Regional fixed effects office	Base model	ME sentiment	Office sentiment	ZGT
Antwerp	1.078***	1.065***	1.126***	1.016***
	[0.152]	[0.110]	[0.386]	[0.139]
Arhus	-0.273	-0.138	-0.115	-0.323*
	[0.196]	[0.146]	[0.433]	[0.179]
Barcelona	-0.484**	-0.781***	-0.478	-0.520***
	[0.213]	[0.162]	[0.403]	[0.196]
Berlin	-1.052***	-1.148***	-0.984**	-1.091***
	[0.184]	[0.134]	[0.387]	[0.165]
Birmingham	-0.13	-0.407**	0.228	-0.151
	[0.237]	[0.172]	[0.520]	[0.214]
Bristol	-0.011	-0.301*	0.202	-0.043
	[0.241]	[0.173]	[0.538]	[0.217]
Brussels	-0.036	-0.023	-0.002	-0.079
	[0.159]	[0.113]	[0.384]	[0.144]
Bucharest	1.462***	1.356***		1.475***
	[0.458]	[0.336]		[0.417]
Budapest	1.246***	0.809***	1.111***	1.232***
	[0.265]	[0.219]	[0.418]	[0.243]
Cardiff	0.31	0.037	0.841	0.262
	[0.262]	[0.192]	[0.659]	[0.238]
Copenhagen	-0.857***	-0.731***	-0.903**	-0.901***
	[0.191]	[0.139]	[0.452]	[0.174]
Cork	1.859***			1.836***
	[0.330]			[0.308]
Dublin	-0.539*		-0.591	-0.566**
	[0.299]		[0.447]	[0.276]
Dusseldorf	-0.911***	-0.951***	-0.814**	-0.929***
	[0.196]	[0.140]	[0.391]	[0.175]
Edinburgh	-0.125	-0.401**	0.096	-0.157
	[0.245]	[0.179]	[0.544]	[0.221]
Frankfurt	-1.012***	-1.086***	-1.045***	-1.057***
	[0.174]	[0.124]	[0.384]	[0.156]
Galway	2.704***			2.674***
	[0.365]			[0.334]
Geneva	-1.915***	-1.878***	-2.222***	-1.956***
	[0.241]	[0.181]	[0.384]	[0.220]
Glasgow	-0.098	-0.356*	0.265	-0.146
	[0.299]	[0.213]	[0.528]	[0.270]
Gothenburg	-0.543***	-0.526***	-0.59	-0.589***
	[0.188]	[0.137]	[0.390]	[0.171]

Note 8.17: The table illustrates the regional fixed effects for the office yield model.

Table 8:18 - Regional fixed effects for the office yield model (2)

Regional fixed effects office	Base model	ME sentiment	Office sentiment	ZGT
Hamburg	-0.795***	-0.872***	-0.738*	-0.826***
	[0.187]	[0.140]	[0.389]	[0.169]
Helsinki	-0.432**	-0.715***		-0.498***
	[0.193]	[0.149]		[0.175]
Istanbul	0.926***	0.448**		0.933***
	[0.197]	[0.181]		[0.171]
Istanbul - Asian CBD	0	0		0
	[0.000]	[0.000]		[0.000]
Istanbul - European CBD	0	0		0
	[0.000]	[0.000]		[0.000]
Krakow	1.238***	0.996***		1.233***
	[0.230]	[0.183]		[0.211]
Leeds	0.013	-0.249	0.423	-0.005
	[0.261]	[0.190]	[0.437]	[0.236]
Liege	0.778	0.881*	1.134***	0.816*
	[0.535]	[0.460]	[0.399]	[0.487]
Limerick	2.570***			2.565***
	[0.792]			[0.727]
London City	-0.743***	-1.032***	-0.497	-0.759***
	[0.258]	[0.186]	[0.401]	[0.232]
London Docklands	0	0	0	0
	[0.000]	[0.000]	[0.000]	[0.000]
London Midtown	-0.696**	-0.980***	-0.385	-0.711***
	[0.296]	[0.212]	[0.425]	[0.266]
London West End	-1.426***	-1.706***	-1.285***	-1.441***
	[0.240]	[0.175]	[0.397]	[0.215]
Luxembourg	-0.15	0.16		-0.176
	[0.206]	[0.148]		[0.187]
Lyon	0.014	0.01	0.024	-0.018
	[0.182]	[0.128]	[0.386]	[0.163]
Madrid	-0.551***	-0.847***	-0.521	-0.577***
	[0.210]	[0.159]	[0.397]	[0.192]
Malmo	-0.292	-0.318***	-0.162	-0.291*
	[0.184]	[0.123]	[0.390]	[0.166]
Manchester	-0.203	-0.467**	0.231	-0.223
	[0.263]	[0.190]	[0.478]	[0.237]
Marseille	0.670***	0.518**	0.577	0.636***
	[0.255]	[0.207]	[0.417]	[0.233]
Milano	-1.124***	-1.322***	-1.216***	-1.135***
	[0.152]	[0.112]	[0.382]	[0.137]

Note 8.18: The table illustrates the regional fixed effects for the office yield model.

Table 8:19 - Regional fixed effects for the office yield model (3)

Regional fixed effects office	Base model	ME sentiment	Office sentiment	ZGT
Moscow	4.189***	3.623***		4.103***
	[0.487]	[0.344]		[0.446]
Munich	-1.389***	-1.457***	-1.379***	-1.420***
	[0.190]	[0.139]	[0.388]	[0.169]
Newcastle	0.204	-0.064	0.382	0.2
	[0.251]	[0.181]	[0.613]	[0.225]
Nottingham	0.34	0.071	0.41	0.314
	[0.238]	[0.180]	[0.758]	[0.216]
Oslo	-0.611**	-0.377**	-0.761*	-0.649***
	[0.244]	[0.172]	[0.424]	[0.220]
Paris (20 districts)	-1.297***	-1.450***	-1.375***	-1.317***
	[0.239]	[0.190]	[0.400]	[0.216]
Paris (CBD)	-1.297***	-1.450***	-1.608***	-1.317***
	[0.239]	[0.190]	[0.410]	[0.216]
Paris Center West included CBD	-1.297***	-1.450***	-1.453***	-1.317***
	[0.239]	[0.190]	[0.410]	[0.216]
Paris Inner Eastern Suburbs	0.261	0.121	-0.063	0.239
	[0.242]	[0.186]	[0.403]	[0.218]
Paris Inner Northern Suburbs	0.024	-0.109	-0.242	0.003
	[0.261]	[0.205]	[0.411]	[0.235]
Paris Inner suburbs (total northern, easthern & southern suburbs)	-0.013	-0.152	-0.26	-0.035
	[0.261]	[0.205]	[0.403]	[0.236]
Paris Inner Southern Suburbs	0.039	-0.118	-0.221	0.018
	[0.269]	[0.208]	[0.407]	[0.243]
Paris Left Bank/Bercy/ Gare de Lyon	-0.554**	-0.682***	-0.735*	-0.571**
	[0.253]	[0.193]	[0.428]	[0.228]
Paris (La Défense)	-0.529**	-0.675***	-0.754*	-0.552**
	[0.237]	[0.177]	[0.402]	[0.214]
Paris Outer suburbs	0.297	0.181	0.409	0.31
	[0.340]	[0.253]	[0.427]	[0.308]
Paris - Western Crescent	-0.749***	-0.743***	-0.809**	-0.764***
	[0.229]	[0.177]	[0.395]	[0.206]
Paris - Western Crescent - Northern Boucle of Seine	0.045	-0.089	-0.046	0.024
	[0.269]	[0.202]	[0.416]	[0.243]
Paris - Western Crescent - Neuilly Levallois	-0.730**	-0.862***	-0.921**	-0.745***
	[0.306]	[0.231]	[0.409]	[0.274]
Paris - Western Crescent - Southern Boucle of Seine	-0.503**	-0.633***	-0.61	-0.522**
	[0.248]	[0.181]	[0.403]	[0.223]
Paris - Western Crescent - Suburbs of La Défense	-0.308	-0.428**	-0.356	-0.325
	[0.284]	[0.212]	[0.419]	[0.257]

Note 8.19: The table illustrates the regional fixed effects for the office yield model.

Table 8:20 - Regional fixed effects for the office yield model (4)

Regional fixed effects office	Base model	ME sentiment	Office sentiment	ZGT
Prague	0.448*	0.657***	0.207	0.405*
	[0.247]	[0.185]	[0.403]	[0.224]
Riga	2.317***	2.373***		2.235***
	[0.376]	[0.282]		[0.344]
Roma	-0.925***	-1.176***	-1.022***	-0.950***
	[0.163]	[0.119]	[0.387]	[0.147]
Rotterdam	0.275	0.253**	0.297	0.246*
	[0.169]	[0.120]	[0.454]	[0.150]
Sheffield	0.745***	0.450**		0.720***
	[0.272]	[0.202]		[0.248]
Stockholm	-1.070***	-1.060***	-0.979**	-1.105***
	[0.184]	[0.133]	[0.396]	[0.167]
The Hague	0.313*	0.303**	0.496	0.277*
	[0.173]	[0.123]	[0.508]	[0.153]
Triangle Area	-0.074	0.131	-0.381	-0.089
	[0.206]	[0.156]	[0.588]	[0.194]
Utrecht	0.247	0.219*	0.328	0.197
	[0.174]	[0.124]	[0.540]	[0.157]
Warsaw	0.419	0.04	0.49	0.362
	[0.282]	[0.208]	[0.705]	[0.257]
Zurich	-1.823***	-1.756***	-2.058***	-1.867***
	[0.173]	[0.132]	[0.387]	[0.158]

Note 8.20: The table illustrates the regional fixed effects for the office yield model.

Table 8:21 - Regional fixed effects for the retail yield model (1)

Regional fixed effects office	Base model	ME sentiment	Retail sentiment	ZGT
Antwerp	0.574**	0.606**	0.575**	0.533**
	[0.269]	[0.252]	[0.287]	[0.240]
Arhus	0.794***	0.896***	0.976***	0.766***
	[0.244]	[0.224]	[0.255]	[0.217]
Barcelona	1.047***	0.877***	1.375***	1.030***
	[0.259]	[0.239]	[0.266]	[0.230]
Berlin	0.406	0.333	0.604**	0.387
	[0.290]	[0.270]	[0.298]	[0.257]
Birmingham	0.333	0.159	2.245***	0.324
	[0.336]	[0.305]	[0.381]	[0.298]
Birstol	0.917***	0.780***	2.888***	0.912***
	[0.277]	[0.256]	[0.348]	[0.246]
Brussels	0.498*	0.529**	0.491*	0.469**
	[0.257]	[0.239]	[0.271]	[0.229]
Bucharest	3.286***	3.315***		3.319***
	[0.553]	[0.515]		[0.493]
Budapest	2.980***	2.308***	3.308***	2.950***
	[0.401]	[0.342]	[0.445]	[0.358]
Cardiff	0.473	0.327	2.444***	0.450*
	[0.302]	[0.281]	[0.432]	[0.269]
Copenhagen	0.182	0.277	0.157	0.152
	[0.297]	[0.275]	[0.306]	[0.267]
Cork	2.538***		3.185***	2.532***
	[0.343]		[0.329]	[0.310]
Dublin	-0.123		0.512	-0.122
	[0.420]		[0.396]	[0.384]
Dusseldorf	0.14	-0.011	0.325	0.136
	[0.346]	[0.295]	[0.358]	[0.307]
Edinburgh	0.444	0.278	2.461***	0.436
	[0.338]	[0.310]	[0.374]	[0.300]
Frankfurt	0.215	0.164	0.385	0.193
	[0.256]	[0.236]	[0.266]	[0.226]
Galway	2.858***		3.523***	2.854***
	[0.712]		[0.692]	[0.651]
Geneva	-0.724***	-0.649***	-0.611**	-0.704***
	[0.242]	[0.221]	[0.249]	[0.220]
Glasgow	0.315	0.143	2.305***	0.288
	[0.341]	[0.309]	[0.378]	[0.303]
Gothenburg	0.798***	0.813***	2.325***	0.770***
	[0.288]	[0.258]	[0.311]	[0.256]

Note 8.21: The table illustrates the regional fixed effects for the retail yield model.

Table 8:22 - Regional fixed effects for the retail yield model (2)

Regional fixed effects office	Base model	ME sentiment	Retail sentiment	ZGT
Hamburg	0.319	0.244	0.510*	0.304
	[0.272]	[0.259]	[0.280]	[0.242]
Helsinki	0.939***	0.769***	1.366***	0.901***
	[0.269]	[0.248]	[0.280]	[0.240]
Istanbul	2.156***	1.967***		2.134***
	[0.470]	[0.490]		[0.419]
Krakow	2.006***	1.883***	2.428***	2.012***
	[0.259]	[0.246]	[0.270]	[0.234]
Leeds	0.772**	0.617**	2.703***	0.765***
	[0.331]	[0.307]	[0.382]	[0.296]
Liege	0.620***	0.621***	0.661***	0.662***
	[0.224]	[0.211]	[0.227]	[0.199]
Limerick	4.086***		4.929***	4.083***
	[0.489]		[0.486]	[0.443]
London West End	-0.4	-0.611**	1.459***	-0.399
	[0.308]	[0.277]	[0.331]	[0.274]
Luxembourg	0.765***	1.034***		0.772***
	[0.261]	[0.250]		[0.233]
Lyon	0.353	0.402	0.374	0.349
	[0.296]	[0.290]	[0.304]	[0.264]
Madrid	0.924***	0.771***	1.247***	0.924***
	[0.255]	[0.238]	[0.260]	[0.228]
Malmo	1.011***	1.064***	2.592***	1.028***
	[0.273]	[0.247]	[0.304]	[0.242]
Manchester	0.453	0.284	2.444***	0.449
	[0.332]	[0.302]	[0.372]	[0.295]
Marseille	1.229***	1.198***	1.329***	1.223***
	[0.307]	[0.309]	[0.315]	[0.277]
Milano	0.743***	0.616**	0.718***	0.739***
	[0.263]	[0.247]	[0.247]	[0.233]
Moscow	7.059***	7.088***		6.970***
	[0.838]	[0.800]		[0.758]
Munich	-0.204	-0.213	-0.058	-0.215
	[0.266]	[0.248]	[0.274]	[0.235]
Newcastle	0.421	0.27	2.355***	0.435
	[0.299]	[0.278]	[0.353]	[0.266]
Nottingham	0.607**	0.442	2.566***	0.596**
	[0.306]	[0.290]	[0.483]	[0.272]
Oslo	0.926***	1.129***	1.395***	0.904***
	[0.327]	[0.298]	[0.336]	[0.290]

Note 8.22: The table illustrates the regional fixed effects for the retail yield model.

Table 8:23 - - Regional fixed effects for the retail yield model (3)

Regional fixed effects office	Base model	ME sentiment	Retail sentiment	ZGT
Paris (20 districts)	0	0	0	0
	[0.000]	[0.000]	[0.000]	[0.000]
Riga	1.757***	1.988***	1.749***	1.754***
	[0.317]	[0.294]	[0.324]	[0.282]
Roma	3.244***	3.414***		3.174***
	[0.461]	[0.426]		[0.416]
Rotterdam	0.809***	0.674***	0.820***	0.795***
	[0.242]	[0.226]	[0.237]	[0.215]
Sheffield	0.393	0.413*	0.466*	0.379*
	[0.257]	[0.237]	[0.266]	[0.230]
Stockholm	0.378	0.424*	1.852***	0.364
	[0.271]	[0.250]	[0.293]	[0.242]
The Hague	0.414	0.454*	0.491*	0.394*
	[0.262]	[0.242]	[0.272]	[0.234]
Triangle Area	0.389	0.568*	0.572	0.411
	[0.336]	[0.314]	[0.350]	[0.312]
Utrecht	0.433*	0.449*	0.495*	0.401*
	[0.263]	[0.244]	[0.274]	[0.236]
Warsaw	1.961***	1.716***	2.054*	1.910***
	[0.365]	[0.323]	[1.163]	[0.327]
Zurich	-0.931***	-0.878***	-0.940***	-0.940***
	[0.275]	[0.280]	[0.295]	[0.247]

Note 8.23: The table illustrates the regional fixed effects for the retail yield model.

Table 8:24 - Regional fixed effects: office yield model (GERUKFRA) (I)

Regional fixed effects office (GERUKFRA)	Base model	ME sentiment	Retail sentiment	ZGT
Birmingham	0.940***	0.584***	1.245***	0.978***
	[0.238]	[0.156]	[0.356]	[0.196]
Bristol	1.058***	0.685***	1.218***	1.074***
	[0.242]	[0.158]	[0.375]	[0.199]
Cardiff	1.379***	1.020***	1.865***	1.355***
	[0.261]	[0.173]	[0.519]	[0.219]
Dusseldorf	0.142	0.195	0.152	0.189
	[0.200]	[0.124]	[0.132]	[0.164]
Edinburgh	0.945***	0.589***	1.109***	0.957***
	[0.243]	[0.163]	[0.387]	[0.202]
Frankfurt	0.04	0.058	-0.101	0.026
	[0.180]	[0.111]	[0.116]	[0.147]
Glasgow	0.972***	0.633***	1.284***	0.947***
	[0.296]	[0.187]	[0.362]	[0.247]
Hamburg	0.257	0.271**	0.246*	0.277*
	[0.193]	[0.126]	[0.127]	[0.160]
Leeds	1.083***	0.739***	1.455***	1.126***
	[0.259]	[0.171]	[0.227]	[0.217]
London City	0.327	-0.041	0.512***	0.376*
	[0.260]	[0.169]	[0.155]	[0.214]
London Docklands	0	0	0	0
	[0.000]	[0.000]	[0.000]	[0.000]
London Midtown	0.374	0.012	0.613***	0.426*
	[0.294]	[0.189]	[0.205]	[0.242]
London West End	-0.356	-0.715***	-0.305**	-0.305
	[0.240]	[0.161]	[0.147]	[0.198]
Lyon	1.068***	1.146***	0.980***	1.084***
	[0.188]	[0.118]	[0.121]	[0.153]
Manchester	0.866***	0.522***	1.238***	0.907***
	[0.263]	[0.171]	[0.293]	[0.217]
Marseilles	1.722***	1.643***	1.530***	1.740***
	[0.258]	[0.188]	[0.196]	[0.217]
Munich	-0.337*	-0.314**	-0.409***	-0.317**
	[0.194]	[0.124]	[0.124]	[0.157]
Newcastle	1.274***	0.926***	1.414***	1.350***
	[0.249]	[0.163]	[0.473]	[0.206]
Nottingham	1.410***	1.061***	1.438**	1.437***
	[0.239]	[0.168]	[0.642]	[0.202]
Paris (20 districts)	-0.244	-0.316*	-0.457***	-0.198
	[0.242]	[0.170]	[0.161]	[0.202]

Note 8.24: The table illustrates the regional fixed effects for the office yield model for the German, French and British city regions.

Table 8:25 - Regional fixed effects: office yield model (GERUKFRA) (II)

Regional fixed effects office (GERUKFRA)	Base model	ME sentiment	Retail sentiment	ZGT
Paris (CBD)	-0.244	-0.316*	-0.705***	-0.198
	[0.242]	[0.170]	[0.183]	[0.202]
Paris Center West included CBD	-0.244	-0.316*	-0.543***	-0.198
	[0.242]	[0.170]	[0.184]	[0.202]
Paris Inner Eastern Suburbs	1.314***	1.252***	0.859***	1.357***
	[0.242]	[0.163]	[0.169]	[0.201]
Paris Inner Northern Suburbs	1.077***	1.024***	0.687***	1.121***
	[0.261]	[0.179]	[0.182]	[0.216]
Paris Inner suburbs (total northern, easthern & southern suburbs)	1.040***	0.979***	0.677***	1.081***
	[0.261]	[0.179]	[0.164]	[0.217]
Paris Inner Southern Suburbs	1.091***	1.014***	0.700***	1.136***
	[0.269]	[0.182]	[0.174]	[0.223]
Paris Left Bank/Bercy/ Gare de Lyon	0.500**	0.456***	0.176	0.551***
	[0.253]	[0.168]	[0.220]	[0.209]
Paris (La Défense)	0.523**	0.459***	0.155	0.564***
	[0.236]	[0.154]	[0.164]	[0.197]
Paris Outer suburbs	1.344***	1.340***	1.383***	1.457***
	[0.335]	[0.216]	[0.211]	[0.280]
Paris - Western Crescent	0.305	0.510***	0.107	0.359*
	[0.230]	[0.161]	[0.147]	[0.190]
Paris - Western Crescent - Northern Boucle of Seine	1.098***	1.044***	0.866***	1.142***
	[0.268]	[0.173]	[0.193]	[0.223]
Paris - Western Crescent - Neuilly Levallois	0.323	0.279	-0.022	0.378
	[0.303]	[0.198]	[0.182]	[0.249]
Paris - Western Crescent - Southern Boucle of Seine	0.550**	0.503***	0.303*	0.598***
	[0.247]	[0.157]	[0.166]	[0.203]
Paris - Western Crescent - Suburbs of La Défense	0.745***	0.710***	0.567***	0.797***
	[0.280]	[0.181]	[0.199]	[0.235]
Sheffield	1.813***	1.433***		1.845***
	[0.269]	[0.183]		[0.226]

Note 8.25: The table illustrates the regional fixed effects for the office yield model for the German, French and British city regions.

Table 8:26 - Regional fixed effects: retail yield model (GERUKFRA)

Regional fixed effects retail (GERUKFRA)	Base model	ME sentiment	Retail sentiment	ZGT
Birmingham	-0.087	-0.304	1.332***	-0.069
	[0.340]	[0.241]	[0.357]	[0.291]
Bristol	0.491*	0.33	1.963***	0.511**
	[0.272]	[0.205]	[0.321]	[0.232]
Cardiff	0.039	-0.124	1.499***	0.025
	[0.300]	[0.230]	[0.395]	[0.260]
Dusseldorf	-0.267	-0.326	-0.279	-0.234
	[0.357]	[0.238]	[0.333]	[0.307]
Edinburgh	0.019	-0.172	1.538***	0.031
	[0.343]	[0.250]	[0.353]	[0.295]
Frankfurt	-0.194	-0.164	-0.214	-0.199
	[0.252]	[0.183]	[0.238]	[0.211]
Glasgow	-0.107	-0.314	1.384***	-0.131
	[0.346]	[0.247]	[0.356]	[0.298]
Hamburg	-0.086	-0.094	-0.095	-0.072
	[0.269]	[0.205]	[0.252]	[0.232]
Leeds	0.351	0.152	1.785***	0.374
	[0.334]	[0.248]	[0.357]	[0.290]
London West End	-0.823***	-1.062***	0.564*	-0.789***
	[0.310]	[0.226]	[0.312]	[0.266]
Lyon	-0.069	0.064	-0.194	-0.045
	[0.300]	[0.231]	[0.287]	[0.258]
Manchester	0.03	-0.172	1.527***	0.055
	[0.336]	[0.242]	[0.348]	[0.289]
Marseilles	0.813***	0.849***	0.748**	0.840***
	[0.306]	[0.260]	[0.294]	[0.265]
Munich	-0.611**	-0.550***	-0.654***	-0.593***
	[0.261]	[0.194]	[0.245]	[0.220]
Newcastle	-0.004	-0.181	1.441***	0.054
	[0.298]	[0.225]	[0.328]	[0.257]
Nottingham	0.176	-0.013	1.626***	0.191
	[0.306]	[0.241]	[0.441]	[0.263]
Paris (20 districts)	0	0	0	0
	[0.000]	[0.000]	[0.000]	[0.000]

Note 8.26: The table illustrates the regional fixed effects for the retail yield model for the German, French and British city regions.

Table 8:27 - Regional fixed effects: office yield model (rEUR) (I)

Regional fixed effects office (rEUR)	Base model	ME sentiment	Office sentiment	ZGT
Antwerp	1.074***	1.075***	1.114***	1.038***
	[0.159]	[0.158]	[0.356]	[0.152]
Arhus	-0.271	-0.272	-0.085	-0.299
	[0.202]	[0.200]	[0.417]	[0.193]
Barcelona	-0.498**	-0.489**	-0.465	-0.518**
	[0.220]	[0.218]	[0.373]	[0.211]
Brussels	-0.043	-0.028	-0.009	-0.068
	[0.165]	[0.164]	[0.353]	[0.157]
Bucharest	1.384***	1.571***		1.392***
	[0.472]	[0.443]		[0.451]
Budapest	1.186***	0.932***	1.095***	1.182***
	[0.272]	[0.290]	[0.391]	[0.260]
Copenhagen	-0.854***	-0.857***	-0.919**	-0.879***
	[0.198]	[0.195]	[0.417]	[0.189]
Cork	1.887***			1.870***
	[0.332]			[0.322]
Dublin	-0.547*		-0.493	-0.564*
	[0.308]		[0.421]	[0.296]
Galway	2.732***			2.710***
	[0.371]			[0.356]
Geneva	-1.885***	-1.878***	-2.142***	-1.909***
	[0.250]	[0.249]	[0.354]	[0.239]
Gothenburg	-0.541***	-0.556***	-0.532	-0.567***
	[0.194]	[0.192]	[0.365]	[0.184]
Helsinki	-0.433**	-0.424**		-0.470**
	[0.198]	[0.196]		[0.189]
Istanbul	0.826***	0.789***		0.835***
	[0.208]	[0.229]		[0.193]
Istanbul - Asian CBD	0	0		0
	[0.000]	[0.000]		[0.000]
Istanbul - European CBD	0	0		0
	[0.000]	[0.000]		[0.000]
Krakow	1.209***	1.225***		1.207***
	[0.241]	[0.237]		[0.231]
Liege	0.809	0.823	1.213***	0.829
	[0.545]	[0.590]	[0.373]	[0.520]
Limerick	2.587***			2.579***
	[0.803]			[0.772]
Luxembourg	-0.141	-0.159		-0.156
	[0.213]	[0.213]		[0.204]

Note 8.27: The table illustrates the regional fixed effects for the office yield model for the remaining European city-regions.

Table 8:28 - Regional fixed effects: office yield model (rEUR) (II)

Regional fixed effects office (rEUR)	Base model	ME sentiment	Office sentiment	ZGT
Madrid	-0.564***	-0.553**	-0.508	-0.579***
	[0.217]	[0.215]	[0.367]	[0.207]
Malmo	-0.288	-0.353**	-0.119	-0.287
	[0.189]	[0.172]	[0.364]	[0.180]
Milano	-1.136***	-1.098***	-1.179***	-1.143***
	[0.158]	[0.155]	[0.351]	[0.150]
Moscow	4.108***	3.947***		4.066***
	[0.499]	[0.431]		[0.478]
Oslo	-0.620**	-0.615**	-0.750*	-0.641***
	[0.252]	[0.250]	[0.391]	[0.240]
Prague	0.445*	0.394	0.199	0.422*
	[0.253]	[0.252]	[0.372]	[0.241]
Riga	2.290***	2.293***		2.245***
	[0.386]	[0.388]		[0.370]
Roma	-0.936***	-0.956***	-0.986***	-0.950***
	[0.169]	[0.167]	[0.357]	[0.160]
Rotterdam	0.276	0.262	0.277	0.26
	[0.174]	[0.172]	[0.421]	[0.164]
Stockholm	-1.068***	-1.084***	-0.922**	-1.087***
	[0.189]	[0.186]	[0.368]	[0.180]
The Hague	0.315*	0.308*	0.498	0.294*
	[0.178]	[0.176]	[0.463]	[0.168]
Triangle Area DK	-0.071	-0.06	-0.384	-0.08
	[0.212]	[0.209]	[0.539]	[0.206]
Utrecht	0.25	0.234	0.33	0.221
	[0.180]	[0.177]	[0.502]	[0.171]
Warsaw	0.393	0.346	0.499	0.362
	[0.290]	[0.279]	[0.652]	[0.276]
Zurich	-1.801***	-1.805***	-2.028***	-1.828***
	[0.179]	[0.179]	[0.358]	[0.171]

Note 8.28: The table illustrates the regional fixed effects for the office yield model for the remaining European city-regions.

Table 8:29 - Regional fixed effects: retail yield model (rEUR) (I)

Regional fixed effects retail (rEUR)	Base model	ME sentiment	Retail sentiment	ZGT
Antwerp	0.574**	0.608**	0.564*	0.541**
	[0.264]	[0.248]	[0.298]	[0.234]
Arhus	0.798***	0.803***	0.934***	0.777***
	[0.239]	[0.219]	[0.265]	[0.212]
Barcelona	1.044***	1.081***	1.540***	1.032***
	[0.253]	[0.236]	[0.278]	[0.225]
Brussels	0.498**	0.531**	0.480*	0.475**
	[0.252]	[0.236]	[0.282]	[0.224]
Bucharest	3.259***	3.376***		3.292***
	[0.539]	[0.522]		[0.479]
Budapest	2.948***	2.368***	3.430***	2.918***
	[0.393]	[0.337]	[0.462]	[0.350]
Copenhagen	0.184	0.181	0.158	0.159
	[0.291]	[0.269]	[0.318]	[0.261]
Cork	2.555***		3.631***	2.553***
	[0.336]		[0.344]	[0.303]
Dublin	-0.116		1.066***	-0.11
	[0.412]		[0.401]	[0.376]
Galwick	2.876***		3.963***	2.877***
	[0.696]		[0.706]	[0.634]
Geneva	-0.701***	-0.650***	-0.460*	-0.681***
	[0.238]	[0.224]	[0.259]	[0.215]
Gothenburg	0.800***	0.788***	3.272***	0.778***
	[0.282]	[0.259]	[0.381]	[0.251]
Helsinki	0.938***	0.981***	1.518***	0.907***
	[0.263]	[0.245]	[0.290]	[0.235]
Istanbul	2.101***	1.888***		2.076***
	[0.458]	[0.438]		[0.409]
Krakow	1.993***	2.033***	2.558***	2.000***
	[0.254]	[0.234]	[0.282]	[0.229]
Liege	0.639***	0.662***	0.637***	0.677***
	[0.222]	[0.209]	[0.235]	[0.197]
Limerick	4.102***		5.331***	4.104***
	[0.480]		[0.515]	[0.435]
Luxembourg	0.771***	0.830***		0.782***
	[0.253]	[0.241]		[0.226]
Madrid	0.923***	0.976***	1.410***	0.926***
	[0.249]	[0.233]	[0.273]	[0.222]
Malmo	1.014***	1.039***	3.525***	1.030***
	[0.268]	[0.251]	[0.378]	[0.237]
Milano	0.741***	0.766***	0.803***	0.738***
	[0.257]	[0.242]	[0.258]	[0.227]

Note 8.29: The table illustrates the regional fixed effects for the retail yield model for the remaining European city-regions.

Table 8:30 - Regional fixed effects: retail yield model (rEUR) (II)

Regional fixed effects Retail (rEUR)	Base model	ME sentiment	Retail sentiment	ZGT
Moscow	7.001***	7.120***		6.917***
	[0.818]	[0.771]		[0.738]
Oslo	0.918***	0.945***	1.619***	0.900***
	[0.318]	[0.297]	[0.352]	[0.281]
Prague	1.759***	1.747***	1.741***	1.759***
	[0.309]	[0.290]	[0.338]	[0.275]
Riga	3.229***	3.284***		3.166***
	[0.450]	[0.419]		[0.405]
Roma	0.806***	0.823***	0.905***	0.794***
	[0.237]	[0.222]	[0.247]	[0.211]
Rotterdam	0.394	0.414*	0.453	0.384*
	[0.252]	[0.235]	[0.276]	[0.225]
Stockholm	0.381	0.399	2.809***	0.371
	[0.265]	[0.247]	[0.363]	[0.236]
The Hague	0.415	0.455*	0.476*	0.400*
	[0.256]	[0.241]	[0.281]	[0.230]
Triangle Area DK	0.405	0.466	0.506	0.43
	[0.329]	[0.312]	[0.360]	[0.305]
Utrecht	0.434*	0.450*	0.482*	0.408*
	[0.258]	[0.241]	[0.284]	[0.232]
Warsaw	1.944***	1.877***	2.162*	1.897***
	[0.356]	[0.319]	[1.197]	[0.319]
Zurich	-0.900***	-0.871***	-0.739**	-0.906***
	[0.273]	[0.259]	[0.306]	[0.245]

Note 8.30: The table illustrates the regional fixed effects for the retail yield model for the remaining European city-regions.

CHAPTER 5 - MACHINE LEARNING APPLICATION

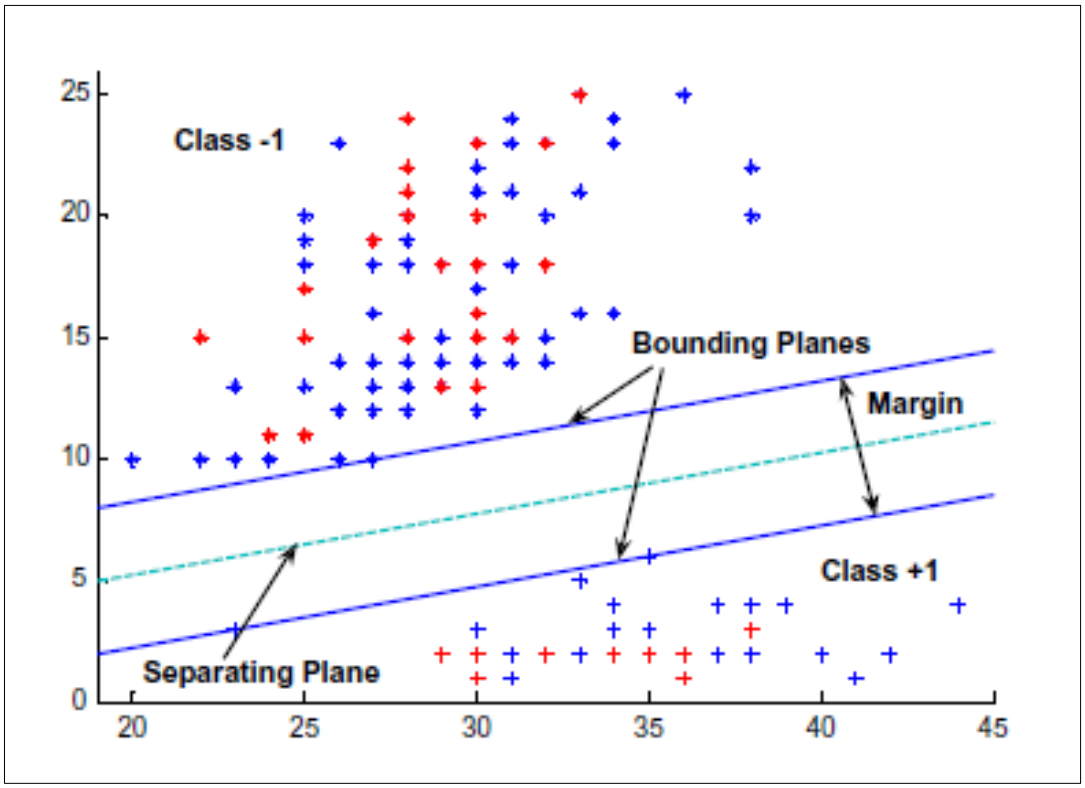
8.1.1 ALGORITHMS

8.1.1.1 SUPPORT VECTOR MACHINE (*SVM*)

Based on the literature *SVM* has been used widely for the classification of text documents [Bai (2011), Yan-Yan et al. (2010), Chen C. C. et al. (2011), Fan et al. (2011), Walker M. A. et al. (2012)]. Nguyen et al. (2015) state that *SVM* is able to handle high dimensional data, which is a good reason why the algorithm is very competitive when it comes to text classification. Medhat et al. (2013) also state that *SVM* is a suitable method for text documents since the sparsity of text allows for a linear classification of the different features. *SVM* belongs to the class of linear classifiers.

In general, the method tries to find the best linear separation between the different classes. This linear separator is called a hyperplane. Initially, *SVM* was applied to binary classification problems, where a linear separation only needed to be achieved between two categories. The method was developed by Vapnik in the 1960s and only many years later published in Cortes and Vapnik (1995). Figure 8:1, taken from Kumar and Gopal (2008), illustrates the original classification issue and the suggested solution.

Figure 8:1 - Geometric interpretation of standard SVM



Note 8.31: The graph illustrates the separation of a dataset by the most optimal hyperplane. The hyperplane tries to maximize the margin between the bounding planes.

The data points are separated by a hyperplane, which tries to find the maximum of the average distance for each of the data points.

In a simplified classification problem with positive and negative data points, we assume that we have a vector $\bar{w}$ of any length which is perpendicular to the median line of the hyperplane (the separating plane in Figure 8:1) and vector $\bar{u}$ which is an unknown data point. We then want to project the unknown in a perpendicular way so that we can figure out on which side of the separating plane the data point lies. This is measured by a constant C.

$$\bar{w} \cdot \bar{u} \geq C \qquad \textit{Equation 8:1}$$

In other words, the dot product of the two vectors plus a constant b ($C = -b$) is assumed to be equal to or larger than 0, which results in the fact that the class is positive.

$$\bar{w} \cdot \bar{u} + b \geq 0, \qquad \textit{Equation 8:2}$$

Equation 8:2 is used as a primary decision rule for further mathematical exploration. Problems are that the constant and $\bar{w}$ remain unknown since not enough constraints have been introduced at this stage. What is known is that beyond the bounding planes the data points will be sorted into either one of the categories, in this simplified case either positive or negative. Using this knowledge, we can transform the unknown vector into a vector $\bar{x_i}$ *or* $\bar{x_j}$ which only represents a clearly classified data point (positive or negative).

$$\bar{u} = \bar{x_i} \qquad \textit{Equation 8:3}$$

Y_i is introduced for mathematical simplification, where Y_i = 1 for a positive sample or $Y_i = -1$ for a negative sample. This results in the equation

$$Y_i(\bar{x_i}\bar{w} + b) - 1 = 0 \qquad \textit{Equation 8:4}$$

for all observations which are directly on the bounding planes. In the case where we would have a unit normal to the width of the hyperplane, which we want to maximize, there is nothing else than

$$width = (\bar{x_i} - \bar{x_j}) \cdot \frac{\bar{w}}{\|\bar{w}\|} \qquad \textit{Equation 8:5}$$

where $\|\bar{w}\|$ represents the magnitude of the vector $\bar{w}$. As a result, we can write

$$MIN\colon \frac{1}{2}\|\bar{w}\|^2 \qquad \textit{Equation 8:6}$$

Which is a result of the decision rule and the planned goal to maximize the hyperplane. The issue here is that we have to address the previously stated constraints in the function where we would like to find the extremes. This can be achieved by using the Lagrange multiplier.

$$L = \frac{1}{2}\|\bar{w}\|^2 - \sum_i \alpha_i \left[Y_i(\bar{w} \cdot \bar{x}_i + b) - 1\right] \qquad \textit{Equation 8:7}$$

After differentiating with respect to a scalar, the vector $\bar{w}$ can be expressed as a linear sum of some of the samples.

$$\bar{w} = \sum_i (\alpha_i \cdot Y_i \cdot \bar{x}_i) \qquad \textit{Equation 8:8}$$

After differentiating Equation 8:7 with respect to the constant b, we achieve

$$\sum_i \alpha_i Y_i = 0 \qquad \textit{Equation 8:9}$$

Now we can combine Equation 8:8 with Equation 8:7

$$L = \frac{1}{2}\left(\sum_i \alpha_i Y_i \bar{x}_i\right) \cdot \left(\sum_j \alpha_j Y_j \bar{x}_j\right) - \sum_i \alpha_i Y_i \bar{x}_i \cdot \left(\sum_j \alpha_j Y_j \bar{x}_j\right) - \sum_i \alpha_i Y_i b + \sum_i \alpha_i \qquad \textit{Equation 8:10}$$

or rewritten

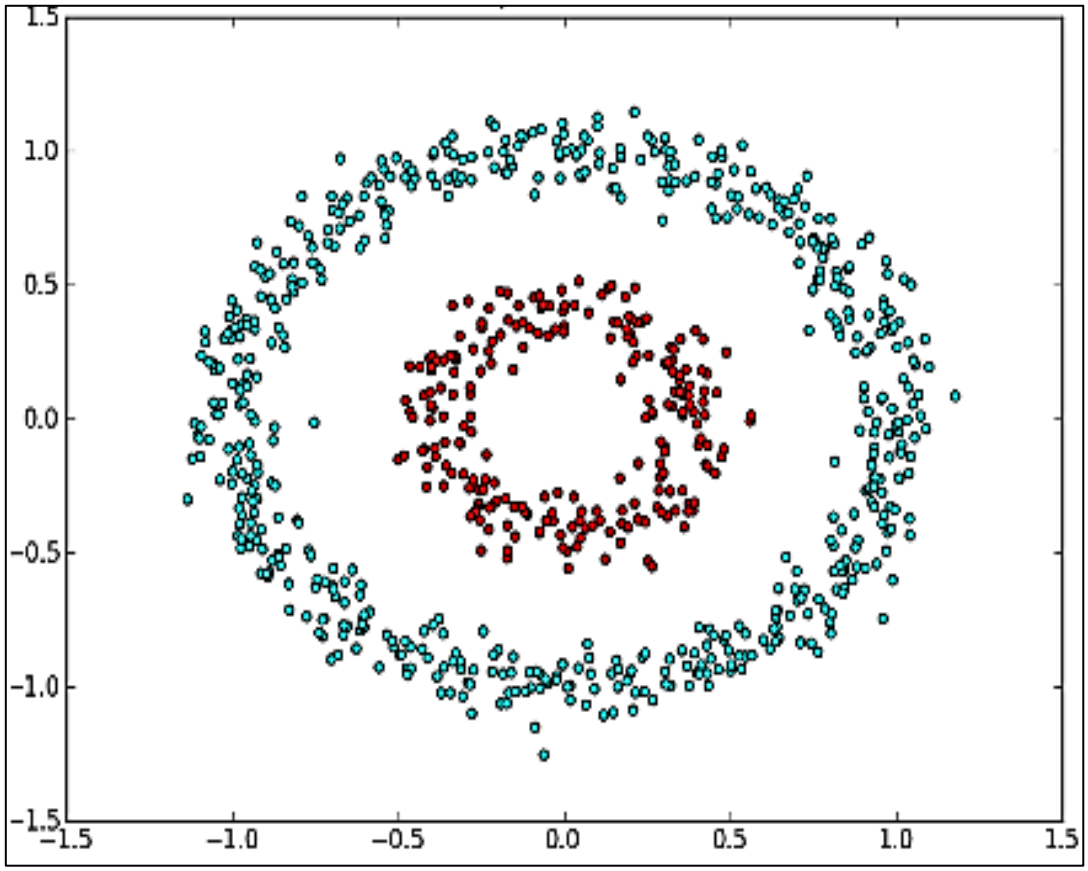
$$L = \sum_i \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \cdot \alpha_j \cdot Y_i \cdot Y_j \cdot \bar{x}_i \cdot \bar{x}_j \qquad \textit{Equation 8:11}$$

Equation 8:11 represents the final equation from which we want to find the extremes. However, it becomes clear that the optimization only depends on the scalar product of the pairs of samples $(\bar{x}_i \cdot \bar{x}_j)$. Going back to the decision rule (Equation 8:2) and replacing the vector $\overline{w}$ with Equation 8:8, we achieve

$$\sum_i \alpha_i Y_i \bar{x}_i \cdot \bar{u} + b \geq 0 \qquad \textit{Equation 8:12}$$

where the optimization depends on $(\bar{x}_i \cdot \bar{u})$. At this stage, it becomes clear that the *SVM* in this form only works in an optimal way, where the classes can be explicitly differentiated. However, in cases where the samples are mixed a linear hyperplane might not be able to separate the data in the most optimal way. Some observations will be unclassified. Figure 8:2 illustrates a case where a linear hyperplane would be unable to sort the data into the correct categories.

Figure 8:2 - Non-linear separable data[36]



Note 8.32: The above-presented figure illustrates a data set, which can not be separated by the application of a linear hyperplane.

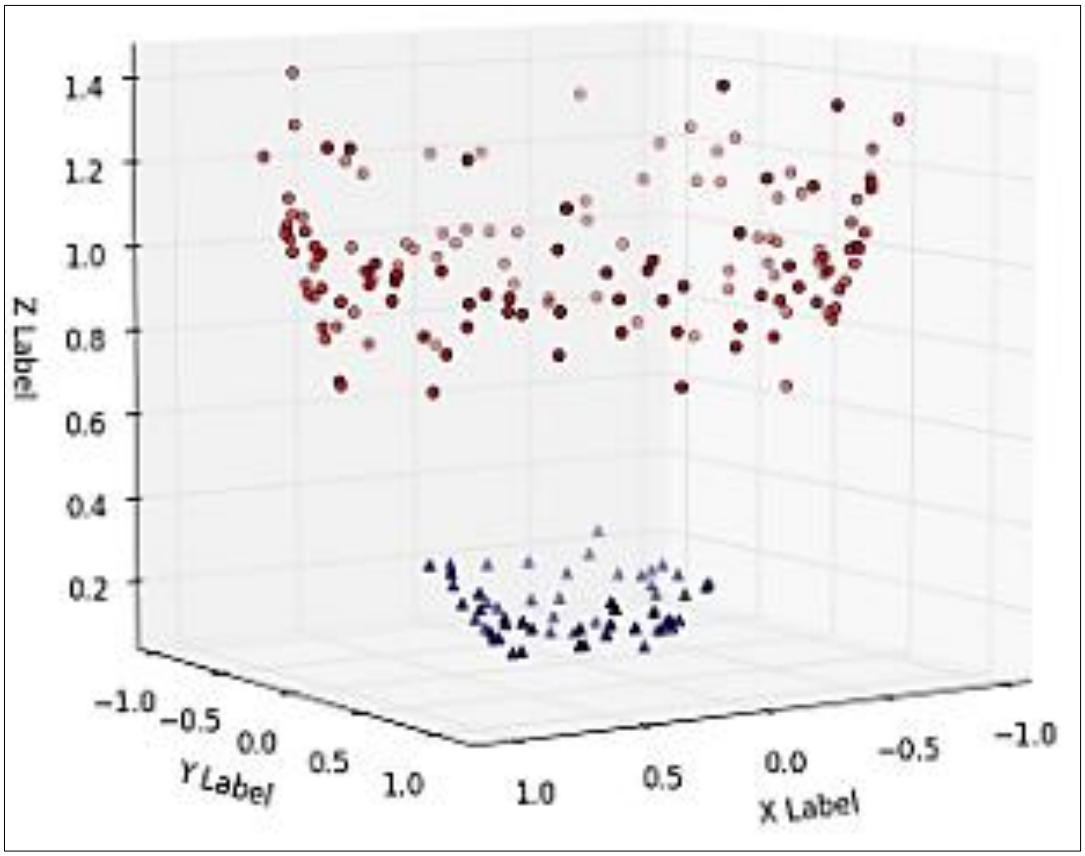
The solution to this issue is the introduction of a different space via the use of a Kernel function $\varphi(\bar{x}_i)$, which we need to maximize.

$$K(\bar{x}_i, \bar{x}_j) = \varphi(\bar{x}_i) \cdot \varphi(\bar{x}_j) \qquad \textit{Equation 8:13}$$

In Equation 8:13 $\bar{x}_j$ can be again replaced with $\bar{u}$. Figure 8:3 shows that a linear solution can be found with the new introduced space.

[36] Graphic from Eric Kim, http://www.eric-kim.net/eric-kim-net/posts/1/kernel_trick.html, accessed on 23 November 2016.

Figure 8:3 - Kernel function applied[37]



Note 8.33: The graph illustrates transformation of the data set from Figure 8:2. Through the application of a Kernel Function the data set has gained a multi-dimensionality. This allows the separation of the data.

In theory, different kernel functions are possible, such as a linear or an exponential kernel.

[37] Graphic from Eric Kim, http://www.eric-kim.net/eric-kim-net/posts/1/kernel_trick.html, accessed on 23 November 2016.

MULTICLASS ISSUE

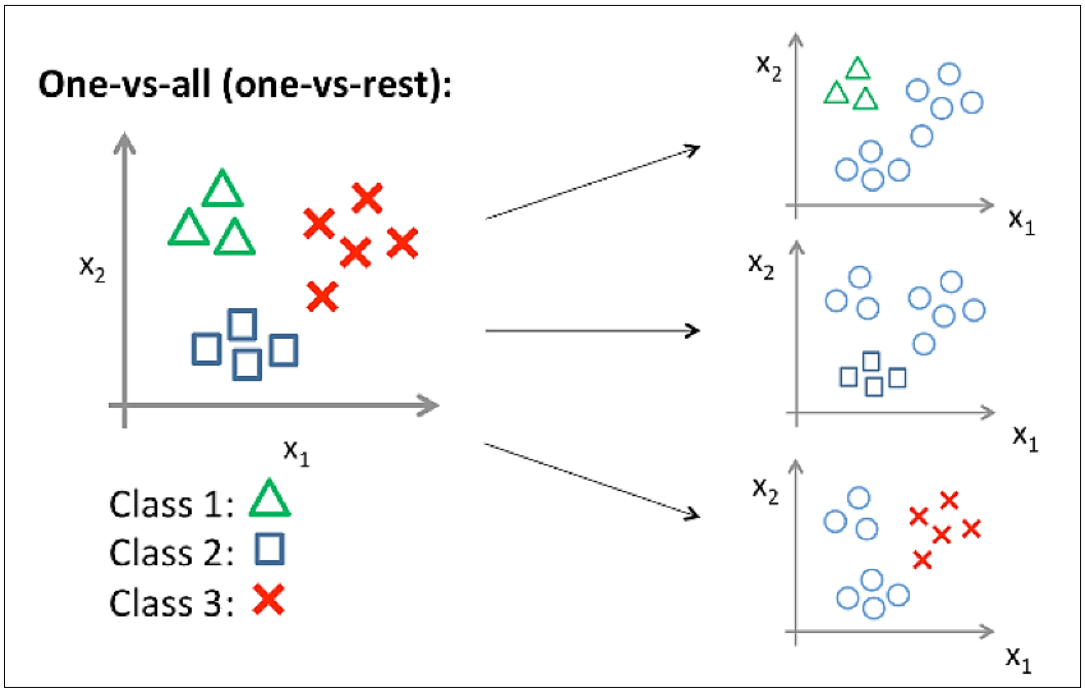
However, the issue which arises based on these mathematical explanations is that the characteristic of the text data is closer to Figure 8:3 and probably even more mixed. Furthermore, the original idea of classifying the news articles based on the star system of *Amazon* (five categories) has not produced any satisfying results.[38] The reasons for this might be that the calculation of this number of options has reached its limits. However, the reduction of classes to three has produced results.[39]

In the literature, the classification of text into more than two categories is described as a multiclass classification issue. The proposed approaches are *one-versus-all* and *one-versus-one*. Hsu and Lin (2002) state that the *one-versus-all* approach calculates *n SVM* models, where *n* represents the number of classes, and then decides for each data point when a maximization has been realized. This assignment is based on probability. This process is computationally expensive since multiple data points are calculated at once for multiple models. Figure 8:4 illustrates the process in more detail.

[38] I stopped the calculation after more than 48 hours, or in other cases the calculation was automatically stopped by the program. The calculation was performed on two different computers: an 8GB and a 128GB ram machine.

[39] The R package does offer for *SVM* the specification of kernel parameters. In this first try I have not applied any specifications and the model has produced results for the three categories. There might be a possibility that the results could be improved by specific kernel arguments.

Figure 8:4 - One-versus-all approach[40]



Note 8.34: The graph illustrates the classification problem with three classes. A linear seperator will separte each class against the other two in order to achieve a clear separation.

On the mathematical side for each of the possible categories, a logistic regression classifier is trained, which is used to predict the probability that an observation can be assigned to a category i.

$$\max_i h_\theta^{(i)}(x) \qquad \textit{Equation 8:14}$$

A new input x will be assigned to a class based on the maximization and its corresponding probability.

The second approach is the *one-versus-one* approach, introduced by Friedman (1996). Here

[40] The figure is taken from https://houxianxu.github.io/2015/04/23/logistic-softmax-regression/, accessed on 24.11.201

$$\frac{k(k-1)}{2}$$ *Equation 8:15*

classifiers are developed, and each classifier is trained on data from two classes.

$$\min_{w^{i,j}b^{i,j}\xi^{i,j}} \frac{1}{2}(w^{i,j})^T w^{i,j} + C\sum_t \xi_t^{i,j}$$ *Equation 8:16*

Equation 8:16 illustrates a binary classification issue which needs to be solved. Friedman (1996) suggests that a voting system for each data point for each class should be applied. After the usage of a kernel function, any x will be sorted based on the suggestion of Equation 8:17.

$$sign((w^{i,j})^T \varphi(x) + b^{i,j})$$ *Equation 8:17*

It seems that the second approach is not as straightforward and that it even takes much more computational power than the one-versus-one approach. However, the *SVM* function in the R - package RTextTools relies on the function in the package e1071 by Meyer et al. (2014). Therefore, the applied code uses the one-versus-one approach with the discussed voting scheme.

8.1.1.2 MAXIMUM ENTROPY CLASSIFIER (*MAXENT*)

The maximum entropy classifier belongs to the class of probabilistic classifiers. A reason for the use of this distribution is that it is uniform. Uniformity equals higher entropy which is desired in this context since no pre-knowledge of the dataset is assumed. A *MAXENT* classifier actually quantifies the uncertainty of the dataset. The entropy of a distribution $H(p)$ is given by the expectation over the surprise

$$H(p) = E_p\left[\log_2 \frac{1}{p_x}\right] = -\sum_x p_x \log_2 p_x \qquad \textit{Equation 8:18}$$

where x is a data point, p_x is the probability and the surprise or uncertainty is given by $\log_2 \frac{1}{p_x}$. It is expected that the distribution maximizes the entropy by minimizing the commitment and that it should be similar to some training data.

Therefore, some constraints are introduced. Every new feature or constraint lowers the maximum entropy and increases the maximum likelihood of the data, and it also transforms the distribution from uniformity towards the actual data. The classifier is actually doing two tasks at the same time. It assigns labels or classes to the test data, and it also estimates a probability distribution over the classifications.

The approach allows for different specifications, which are based on the data and our expectations. In a case where no constraints are introduced the classifier assigns to each event the same probability. If there is pre-knowledge of the data and its distribution, then we could assign different expected distributions to each micro-stage. To summarize, the best model created by a *MAXENT* classifier is the one which allows for the most uncertainty from the data.

The *MAXENT* classifier has been used for text classification. In Nigam et al. (1999) the application is discussed in further detail. The authors state right at the beginning that the performance of the classifier is influenced mainly by the text corpus. In experiments on different corpora, the classifier has performed both better and worse in comparison to the Naive Bayes classifier. Using *MAXENT* in a supervised learning fashion, the constraints for the classifier are introduced by the training dataset. This shows that the training data and the test data should have a common ground, in other words, if they do not match in their topic or origin, the test data will not be classified in the best way. Based on the training data each real-valued function

of the document and the class is set as a feature for the test data. The learned conditional probability distribution is given by

$$\frac{1}{|D|}\sum_{d\in D} f_i\big(d, c(d)\big) = \sum_{d} P(d)\sum_{c} P(c|d)\, f_i(d,c) \qquad \textit{Equation 8:19}$$

where D is the training data, $f_i(d,c)$ is a feature, $P(c|d)$ represents the conditional distribution and $P(d)$ is the document specific distribution. The latter one is unknown and the training data is used for the estimation after considering the constraints

$$\frac{1}{|D|}\sum_{d\in D} f_i\big(d, c(d)\big) = \frac{1}{|D|}\sum_{d\in D}\sum_{c} P(c|d)\, f_i(d,c) \qquad \textit{Equation 8:20}$$

In this study, the constraints are the different classes, which will be estimated based on the training dataset.

The *MAXENT* classifier carries the risk of overfitting, which could be overcome by introducing different priors. In this study, I have opted not to introduce any priors and other constraints, since everything is unknown in the two datasets, except the distribution of the classes.

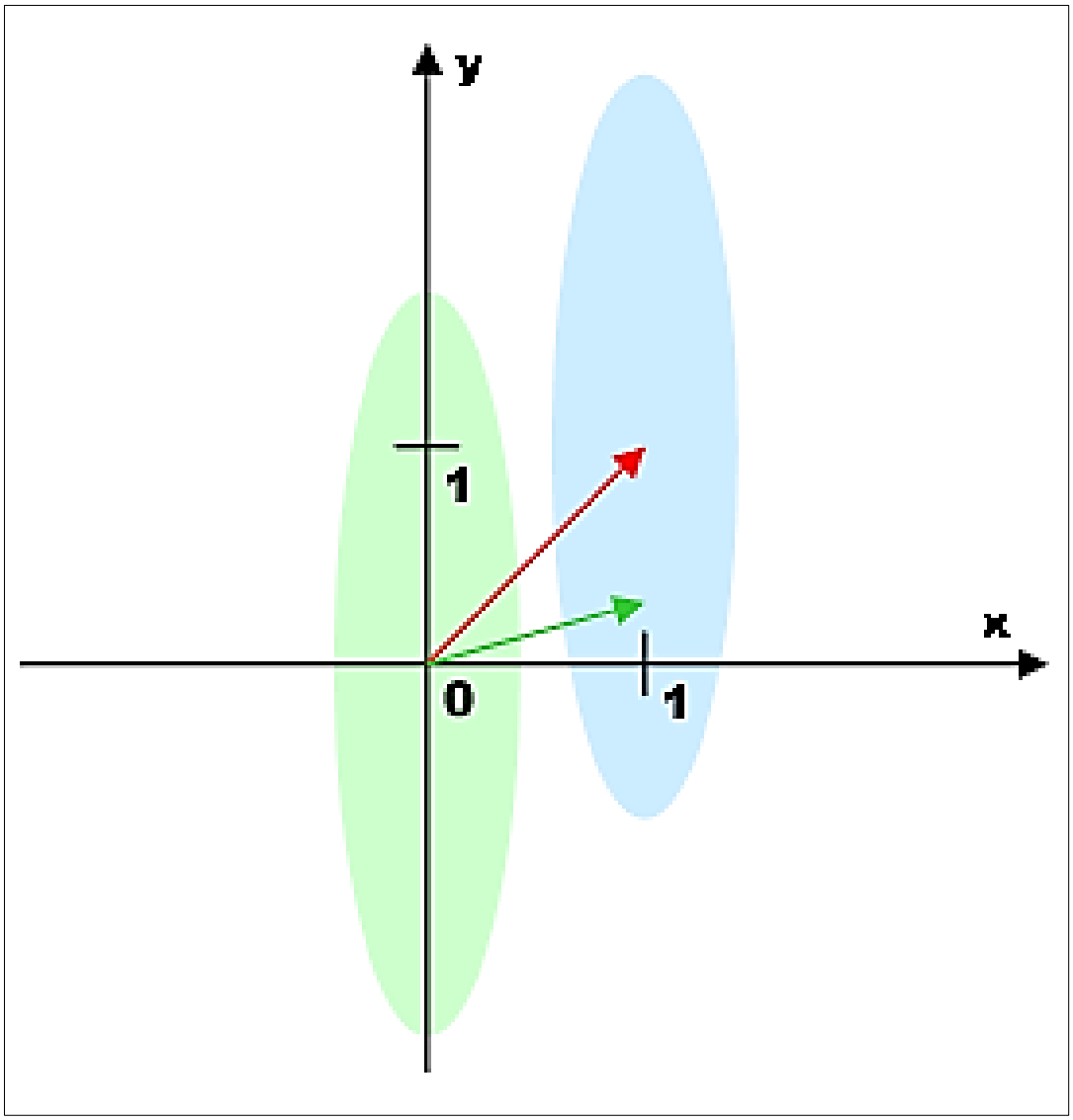
I am aware of the fact that extended work can be performed on the corpora to improve these results.

8.1.1.3 STABILIZED LINEAR DISCRIMINANT ANALYSIS (SLDA)

The SLDA approach has not been widely applied to text classification in comparison to other classifiers. It does further seem that the authors of the package have mixed up the names of the approach. In Jurka et al. (2013), SLDA is stated as Scaled Linear Discriminant Analysis with reference to the ipred package of Peters et al. (2013), who themselves state SLDA as Stabilized Linear Discriminant Analysis. I will follow the latter definition in this study.

LDA belongs to the class of linear classifiers and generalizes Fisher's linear discriminant. The method is similar to the support vector machine technique. LDA tries to separate two or more classes with a linear classifier. The original LDA proposed by Fisher (1948) shows similarities to regression analysis and other separating statistical methods such as principal component analysis (PCA) or factor analysis. In comparison to PCA, LDA considers differences between the classes in the estimation process to guarantee a maximum of separation. The process of PCA changes the location and the shape of the original data, which remains untouched by LDA. Figure 8:5 illustrates the problem set and the suggested solution by Fisher.

Figure 8:5 - Application of the Fisher LDA[41]



Note 8.35: The graph illustrates the LDA process. Two goals are tried to achieve. First the dimensionality is reduced and second the reduction should also provide a reasonable separtion of the two datasets in order to avoid overfitting. Since the process is comparable to a PCA, both methods try to find a new common component. However, the added advantage of LDA is to tackle overfitting.

The figure shows two classes which are centred around the points $(0,0)$ and $(1,1)$. The most natural solution would be a straight line between the two points (red arrow) and project all other observations on it. However, due to the fact that the classes should overlap this is not feasible. Fisher suggested finding another axis which maximizes the below stated $J(w)$.

Two classification approaches are common with LDA, a class-dependent and a class-independent transformation. In the first case, the maximization is reached by focusing on the

[41] Figure taken from http://www.alglib.net/dataanalysis/lineardiscriminantanalysis.php, accessed on 29 November 2016.

within-class variance, where with the second approach the maximization is attempted at an overall level. Further, the two approaches differ in the number of criteria they need for the process.

Again, starting with the case where the data is sorted into two different classes, LDA uses the given observations $\vec{x}$ of the training data with the observed classes y. The algorithm assumes a normal distribution with

$$p(\vec{x}|y) = 0 \quad and \quad p(\vec{x}|y) = 1 \qquad \textit{Equation 8:21}$$

and a mean of μ. The means of the two classes in the training dataset are given by μ_1 and μ_2.

$$\mu_3 = p_1 * \mu_1 + p_2 * y_2 \qquad \textit{Equation 8:22}$$

This results in the overall mean μ_3, given by the probabilities p_n of the corresponding class. Welling (2005) states that the between-class S_B and the within-class S_w scatter matrix is used to achieve the separation. They are defined as:

$$S_B = \sum_c (\mu_c - \bar{x})\,(\mu_c - \bar{x})^T \qquad \textit{Equation 8:23}$$

$$S_W = \sum_c \sum_{i \in c} (x_i - \mu_c)\,(x_i - \mu_c)^T \qquad \textit{Equation 8:24}$$

Based on this the general transformation rule for scatter matrices can be applied to estimate a new vector.

$$S_{\mu+v} = S_{\mu} + N_{vv}^{T} + 2N_{v}(\mu - \bar{x})^{T} \qquad \textit{Equation 8:25}$$

Equation 8:23 and Equation 8:24 can be ultimately used to represent Fisher's linear discriminant.

$$J(w) = \frac{w^{T} S_{B} w}{w^{T} S_{W} w} \qquad \textit{Equation 8:26}$$

$J(w)$ represents the ratio of the total sample variance to the sum of variances within the separate classes.

In Brenning (2009) it is stated that SLDA is able to handle high-dimensional data. The stabilization of the classifier according to Läuter (1992) is realized by reducing the dimension of the feature space, which leads to a digital stabilization of the classifier.

Again, it is fair to mention that LDA or SLDA have not been widely used for the task of text classification. Other fields where the algorithm has been applied are speech recognition, face recognition and biomedical studies [David et al. (2010)].

8.1.1.4 LASSO AND ELASTIC-NET GENERALIZED LINEAR MODELS (*GLMENT*)

The *GLMENT* method which is used in the RTextTools R - package is based on the same method as in the *GLMENT* R - package by Friedman et al. (2009). In Friedman et al. (2010) the authors specified in more detail their application. The algorithm was developed for the estimation of generalized linear models with convex penalties. Different regression methods are covered, and three penalties (ℓ) are applied, such as the lasso, the rigid regression or a combination of the two – an elastic net. Friedman et al. (2010) state that in general a cyclical coordinate descent with computations around the regularization path is applied and that *GLMENT* performs well with significant problems with a high number of variables. However, Medhat et al. (2013) have not recorded any study where the algorithm has been applied to text classification. According to Hastie and Qian (2014), the algorithm also fits logistic, nominal, Poisson and Cox regression models, as well as multi-response regression models.

The application tries to solve:

$$\min_{\beta_0,\beta} \frac{1}{N} \sum_{i=1}^{N} w_i l(y_i, \beta_0 + \beta^T x_i) + \lambda P_\alpha(\beta) \qquad \textit{Equation 8:27}$$

and

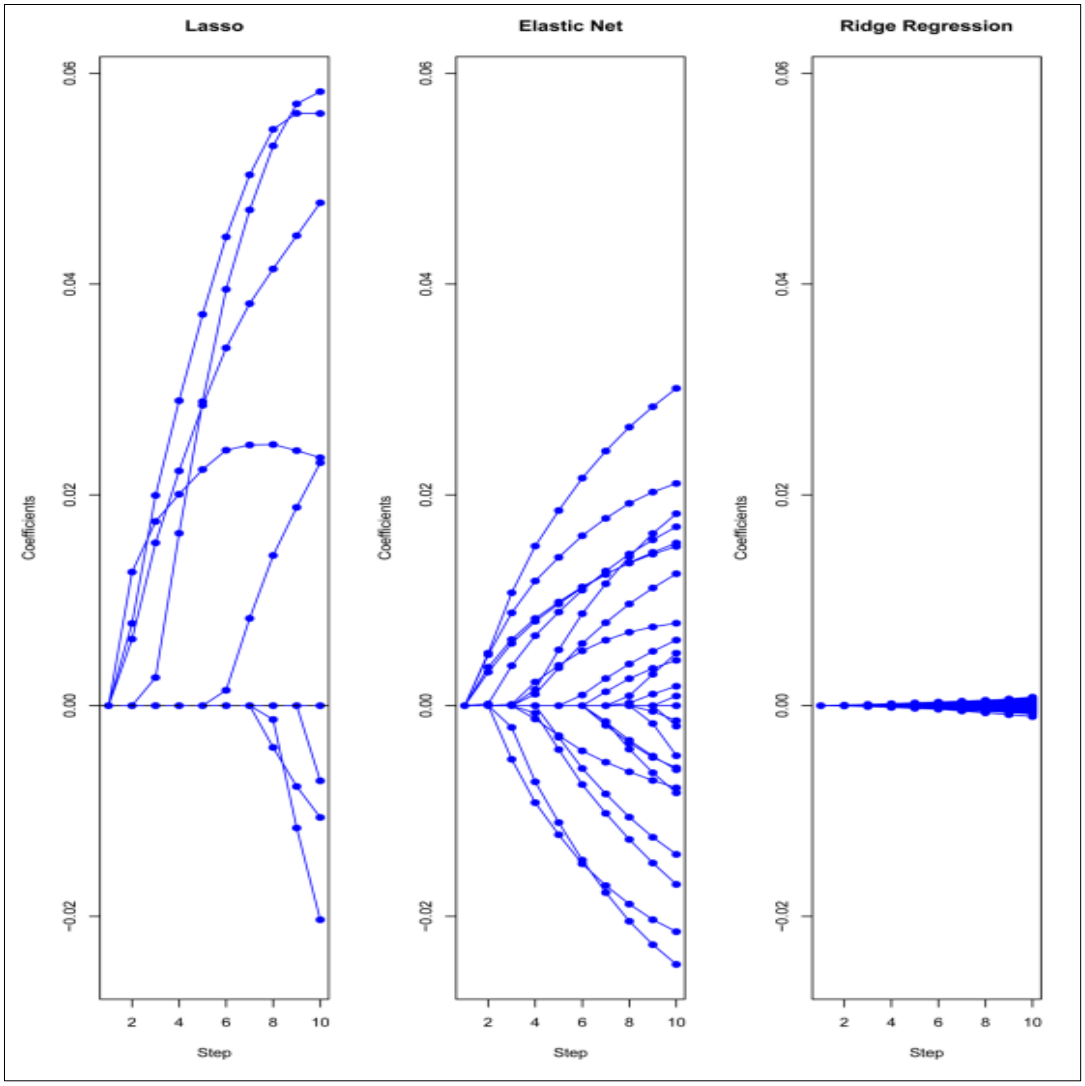
$$P_\alpha(\beta) = (1-\alpha)\frac{1}{2}\left|\left|\beta\right|\right|_{\ell_2}^2 + \alpha||\beta||_{\ell_1} \qquad \textit{Equation 8:28}$$

where the values of λ (from max to min) cover the entire range. The negative log-likelihood given by $l(y, \eta)$ contributes to the observations i. α represents the elastic-net penalty, where P_α bridges the two penalties lasso and rigid. If the default function were to use $\alpha = 1$, the lasso could take the value of 0 for the rigid regression. The penalty therefore depends on the value of α and leaves room for interpretation.

Both the lasso and the rigid penalties have their drawbacks, which is solved by the elastic net. According to Friedman et al. (2010), the stiff penalty tends to shrink the coefficients of correlated predictors towards each other to gain extra explanatory power. If there are identical predictors, they end up having the same coefficient. Lasso instead selects one predictor over

the other. This approach is orientated on a Laplace prior, where many coefficients are assumed to be close to zero and a minority is more substantial. α further provides numerical stability and if corrected it can work as a lasso and removes any extremes caused by high correlations in the elastic net framework $\alpha = 1 - \epsilon$, with $\epsilon > 0$. Figure 8:6 illustrates the mechanics of the three measures applied to leukaemia data, where for the elastic net $\alpha = 1 - 0.8$ has been used.

Figure 8:6 - Example of the different penalties[42]



Note 8.36: The figure above compares the three different penalties: Lasso, Elastic Net and rigide regression. Both the lasso and rigide regression will push the results to a more extreme outcome. The Lasso or the Least Absolute Shrinkage and Selection Operator, uses a penealty term, which shrinks the regression coefficients toward zero. The term is the sum of the absolute coefficients. The Ridge regression on the other hand, shrinks the regression coefficients of variables with minor contribution to the outcome. They are set close to zero. The Elastic Net approach combines both methods and penalizeses with both penalties at the same time. Therefore, the coefficients, were appropriate are either shrinked (ridge regression) or set close to zero (LASSO).

[42] Graph taken from Friedman et al. (2010).

It can be seen that the lasso and the rigid approach are more extreme in their estimations, where the elastic net tries to find a middle ground.

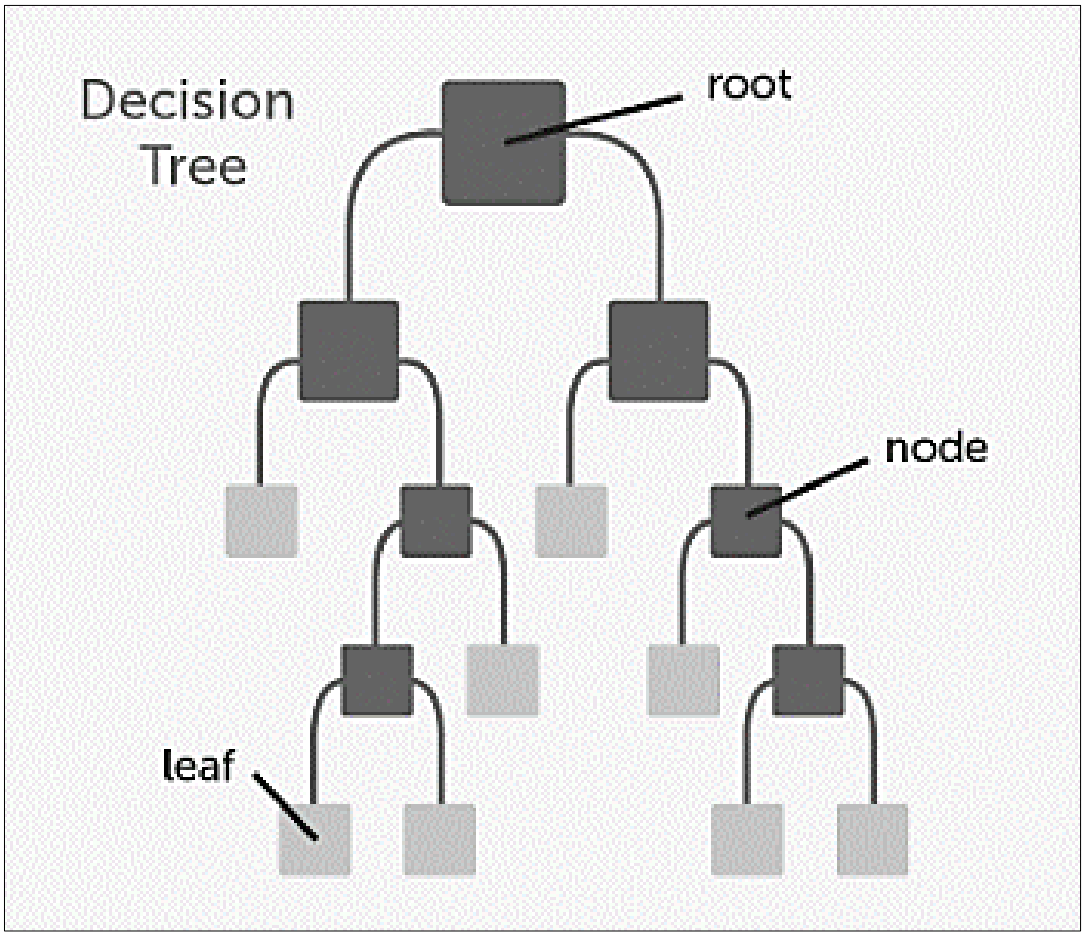
Further, the model, as well as the package, next to the adjustment for the α value, allows for further modifications. These depend on the selected model.

8.1.1.5 DECISION *TREE*

For the following methods, the decision *TREE* is used as a structural base. Different to other approaches decision *TREE*s are easy to understand, interpretable and controlled [Ertel (2011)] since they allow us to observe how a specific observation x is actually classified. Another advantage is that problem sets can be directly sorted into multiple classes.

In general, the algorithm is a top-down method with the root node at the top and with different nodes attached to it; the lowest levels are the leaves, which can be seen as the classes or labels. During the classification process, some leaves can remain empty. One main issue is that it is necessary to control the growth of the *TREE* by selecting good splits and by deciding when a sufficient number of levels has been reached [Breiman et al. (1984)].

Figure 8:7 - Structure of a decision TREE[43]



Note 8.37: The figure above illustrates the process of a decission tree. The entity will be pushed through all decision nods until it has reached one of the final leafs. Each leaf can be compared to a specific category.

At each node, the observation is compared to some criteria and then sent to either one of the directions based on the information content. The observation always follows the path with the highest information. This is also called binary separation, but it is a problem since each split must be able to separate the data into smaller classes. If the splits are not efficient enough, then the classification process will be disturbed. Similar to the *MAXENT* approach the decision *TREE* relies on entropy $H(p)$ as a measure of information content. Equation 8:18 has illustrated the calculation of entropy. Following this definition then, an event with no uncertainty $p = (1, 0, \dots, 0)$ would solve the equation

[43] Figure taken from http://www.aunalytics.com/decision-trees-an-overview/, accessed on 6 December 2016.

$$H(p) = -\sum_{i=1}^{n} 0 \log_2 0 = 0$$ *Equation 8:29*

Since each of the datasets has an assigned probability p, the concept of entropy can be extended to the data D. The decision *TREE* starts with all the training data in the top node and eventually partitions the set down to the leaves. This recursive partitioning should create classes with a pure character so that the label is unique.

$$H(D) = H(p)$$ *Equation 8:30*

With the decision *TREE* the uncertainty should be reduced, and therefore the information content $I(D)$ will be maximized

$$I(D) \coloneqq 1 - H(D)$$ *Equation 8:31*

The structure of the *TREE* with its different nodes divides the data on each node into smaller subsets. Each node can be seen as a question or attribute against which an observation is compared. The smaller the remaining dataset is, the better is the separating node. The information gain $G(D, A)$ is defined by

$$G(D, A) = \sum_{i=1}^{n} \frac{|D_i|}{|D|} I(D_i) - I(D)$$ *Equation 8:32*

This results in the decision rule for each of the individual nodes.

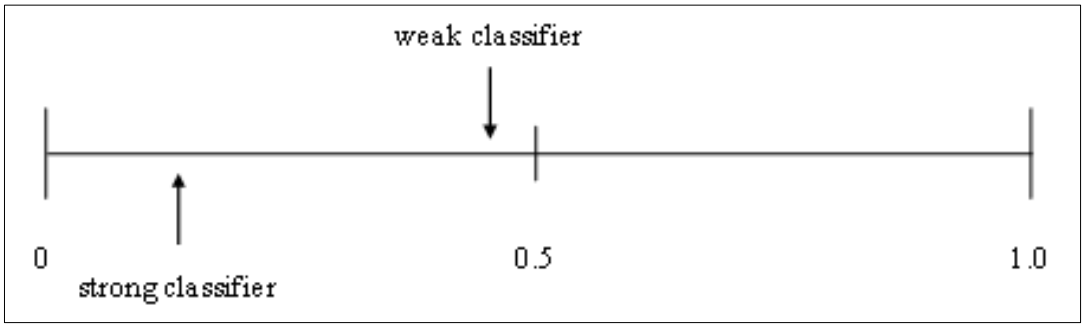
$$G(D,A) = H(D) - \sum_{i=1}^{n} \frac{|D_i|}{|D|} H(D_i)$$ *Equation 8:33*

The applied algorithm relies on the *TREE* package by Ripley (2007). Unfortunately, the algorithm was producing unsatisfying results in this study. A reason for this can be seen in the data. Due to the hierarchical structure of the decision *TREE*, the training dataset is further and further decomposed until a minimum number of instances is collected in a leaf. The issue with the text data is that the separation is based on specific words, whether they are present or not. However, as shown above, the text distributed over the different classes shows some similarity. It seems that the nodes or the attributes at each node were not strong enough to separate.

8.1.1.6 BOOSTING

BOOSTING is not a stand-alone algorithm as *SVM* or *MAXENT* is. The process of *BOOSTING* somewhat describes a specific method where multiple algorithms are used to solve a classification problem. In other words, it depends on the wisdom of the crowd. Starting with the assumption that a weak learning algorithm exists, which just performs slightly better than a random classifier, *BOOSTING* tries to improve this algorithm (Figure 8:8).

Figure 8:8 - Classification categories based on their error rate



Note 8.38: The graph above illustrates the categorization of the classifiers. The lower the error rate of an classifier the better it is. Classifiers which reach an error rate of 50% can be compared to a random guessing process.

The improvement is reached by continually drawing back to this existing classifier and the training data. In Schapire and Freund (2012) the authors state that even weak classifiers have benefited since their error rate is slightly better than a random classifier or a guessing approach; this is the central idea of *BOOSTING*. A random classifier would be a coin flip, with a 50% chance of predicting the outcome of the next coin flip correctly. In general, the approach uses a voting system among the different classifiers.

Like the previous examples *BOOSTING* dealt initially with binary classification issues, given a training dataset with $(x_1, y_1), \dots, (x_m, y_m)$ with x_i instances and y_i corresponding labels. The labels are either $+1$ or -1. Since the base model will only produce weak results the training data needs to be modified to achieve better results.

$$H(x) = sign\left(h^1(x) + h^2(x) + \dots + h^m(x)\right) \qquad \textit{Equation 8:34}$$

Equation 8:34 illustrates the applied method. The *BOOSTING* algorithm $H(x)$ relies on several algorithms, where only the sign of the equation is of interest. If the majority of algorithms produce the correct result, then the sign will be correct.

A new classifier will only choose a sample of the training data, where the base model has significantly underperformed. The algorithm, therefore, runs multiple iterations to improve the overall result.

$$Data \rightarrow h^1$$
$$Data_{exaggeration\ of\ h^1 errors} \rightarrow h^2$$
$$Data_{exaggeration\ of\ h^{m-1} errors} \rightarrow h^m$$

Equation 8:35

For each chosen sample from the training dataset a distribution D_t is maintained; each of these sub-samples is given a specific weight w_i. Each weight provides information about the correctly specified instances of the corresponding classifier and can be used as a measure. At the beginning of each iteration these weights are equal; however, they shift towards more difficult samples, where the algorithm needs to invest more time for the solution. The errors are calculated as in Equation 8:36, where N is the number of samples; with the basic assumptions that the weights are equally distributed.

$$\varepsilon = \sum_{wrong} \frac{1}{N}$$

Equation 8:36

$$w_i^1 = \frac{1}{N}$$

Equation 8:37

$$\varepsilon = \sum_{wrong} w_i$$

Equation 8:38

with the overall distribution,

$$\sum w_i = 1 \qquad \textit{Equation 8:39}$$

Therefore, Equation 8:34 can be rewritten by considering the weights,

$$H(x) = sign\left(\alpha^1 h^1(x) + \alpha^2 h^2(x) + \ldots + \alpha^m h^m(x)\right) \qquad \textit{Equation 8:40}$$

From here only the classifier h^t is chosen which minimizes the ε^t errors at time t, to compute α^t. This classifier will predict w^{t+1} and will be updated in a loop until a satisfactory result for alpha has been found.

$$w_i^{t+1} = \frac{w_i^t}{Z} e^{-\alpha^t h^t(x) y(x)} \qquad \textit{Equation 8:41}$$

Here Z represents a normalization factor, which secures a new combination of weights that adds up to one. $y(x)$ is a function which is either +1 or –1, depending on expectations. The minimum error bound can be found, if

$$\alpha^t = \frac{1}{2} \ln \frac{1 - \varepsilon^t}{\varepsilon^t} \qquad \textit{Equation 8:42}$$

This results in,

$$w_i^{t+1} = \frac{w_i^t}{Z} * \begin{cases} \sqrt{\frac{\varepsilon^t}{1-\varepsilon^t}} & correct\ prediction \\ \sqrt{\frac{1-\varepsilon^t}{\varepsilon^t}} & wrong\ prediction \end{cases}$$

Equation 8:43

The normalization factor is defined by

$$Z = \sqrt{\frac{\varepsilon^t}{1-\varepsilon^t}} \sum_{correct} w_i^t + \sqrt{\frac{1-\varepsilon^t}{\varepsilon^t}} \sum_{wrong} w_i^t$$

Equation 8:44

$$Z = 2\sqrt{\varepsilon^t(1-\varepsilon)}$$

This finally results in

$$w_i^{t+1} = \begin{cases} \frac{w^t}{2} * \frac{1}{(1-\varepsilon)} & correct \\ \frac{w^t}{2} * \frac{1}{\varepsilon} & wrong \end{cases}$$

Equation 8:45

and

$$\sum_{correct} w_i^{t+1} = \frac{1}{2} \quad and \quad \sum_{wrong} w_i^{t+1} = \frac{1}{2}$$

Equation 8:46

The sum of these weights is a scaled version of their previous version.

From the original classification issue, we can summarize that not all applied tests are necessary. Those tests which are performed between two correctly specified classifiers are needless. Therefore, only a small number of tests is required. The advantages of this method can be found in the fact that the algorithm does not overfit, such as happens in other approaches like *SVM* or *MAXENT*. The reasons for this phenomenon remain unclear.

Nevertheless, this method needs to be adjusted for a multiclass problem $K > 2$. The main issue is that the approach is based on the binary classification. One way would be the one-against-all approach where a range of yes or no questions will be asked; this however might result in an unnecessary amount of calculations. Following Schapire and Freund (2012), this adjustment is reached by

$$H(x) = \ arg \max_{y \in Y} \sum_{t=1}^{T} \alpha_t 1\{h_t(x) = y\} \qquad \textit{Equation 8:47}$$

Yet, the problem arises regarding the initially established weight of the error. In the case of a random guess with a binary issue, this would result in $\frac{1}{2}$. The above-stated method assures that ε will be below this value, so that the error for the combined analysis decreases dramatically. This cannot be realized with multiple classes since the minimal error distribution would be $\frac{1}{K}$. So, the basic requirement would be further emphasized, namely that the basic classifier needs to be better than 50%. In the binary case, a weak classifier which is worse than this hurdle is simply replaced by its negation, $-h_t$. This, however, cannot be done in a multiclass issue. Therefore, the performance of the initial classifier is of tremendous importance. In the case where it already produces a higher error rate, it would result in no improvement. Unfortunately, the applied algorithm just stops and accepts the poor initial result.

The used function in the code relies on decision *TREE* stumps. Different to the *TREE* structure where multiple branches exist, here the root node is directly linked to the leaf. These stumps are also called one-level decision *TREE*s [Iba and Langley (1992)] and are specified as weak learners.

8.1.1.7 BAGGING: BOOTSTRAP AGGREGATION

BAGGING is modifying the previously shown method of *BOOSTING*. The idea is that a range of different classifiers is used to improve a base classifier. However, different to *BOOSTING*, where the majority vote of the different classifiers h^n is used to label an observation x, which could result in an increase of the expected classification error, BAGGING uses bootstrapped samples from the original dataset and the samples are adjusted for each iteration. Sometimes BAGGING is also called "bootstrap aggregating", which underlines this difference to *BOOSTING*. The distribution D_t is fixed so that each iteration remains uniform over the training data.

With each iteration, the base classifier is trained on a bootstrapped sample. Some of the observations are more influential than others since they will be selected more often. According to Schapire and Freund (2012), one-third of all observations will be omitted on average. Further, following the authors, the advantage of BAGGING can be seen in the fact that it is successful in handling data with significant variance. In this framework, the variance has been defined as the amount of decrease in the error affected by BAGGING. Theoretically, each bootstrapped sample should approximate a genuinely independent sample. Nevertheless, it comes down again to the original base classifier: if this one is already dominated by variance, then the resulting classification suffers.

For a more formal description of the algorithm, I use the mathematical explanation of Breiman (1996), where it is assumed that y the class and x the observations in $\mathcal{L}$, the test dataset, are taken from the probability distribution P, therefore an aggregated predictor is defined as

$$\phi_A(x) = E_{\mathcal{L}}\phi(x, \mathcal{L}) \qquad \textit{Equation 8:48}$$

Using the observations to generate the classes,

$$E_{\mathcal{L}}(y - \phi(x, \mathcal{L}))^2 = y^2 - 2yE_{\mathcal{L}}\phi(x, \mathcal{L}) + E_{\mathcal{L}}\phi^2(x, \mathcal{L}) \qquad \textit{Equation 8:49}$$

This results, after using Equation 8:48 to modify Equation 8:49 with respect to inequality $EZ^2 \geq (EZ)^2$, in

$$E_{\mathcal{L}}(y - \phi(x, \mathcal{L}))^2 \geq (y - \phi(x, \mathcal{L}))^2 \qquad \textit{Equation 8:50}$$

Over the joint distribution of x and y, the mean squared error of $\phi_A(x)$ will be lower than the averaged mean squared error of $\phi(x, \mathcal{L})$; this depends on the size of the inequality of the two sides.

$$[E_{\mathcal{L}}\phi(x, \mathcal{L})]^2 \leq E_{\mathcal{L}}\phi^2(x, \mathcal{L}) \qquad \textit{Equation 8:51}$$

The problem with Equation 8:48 is that improvement can only be achieved if the two sides differ; however, if they are similar, then no improvement will be achieved. Therefore, $\phi(x, \mathcal{L})$ is preferred to be variable. Yet, ϕ_A is always improving upon on ϕ.

Considering the probability distribution over $\mathcal{L}$, ϕ_A depends on both x and P, the bagged estimator is given by

$$\phi_B = \phi_A(x, P_{\mathcal{L}}) \qquad \textit{Equation 8:52}$$

where $P_{\mathcal{L}}$ is the bootstrapped estimation of P. ϕ_B which is also influenced by the stability of the process. In the case of an unstable process, improvement is achieved by aggregation, where in the case of a stable process ϕ_B accuracy suffers. This can lead to the case where ϕ_B damages the classification process instead of improving it. Similar to *BOOSTING*, it might also be the case that the base classifier is near maximum accuracy, which results in no further improvement through BAGGING.

The defined classifier $\phi(x, \mathcal{L})$ is then used to predict a feature or class $j \in \{1, \dots, J\}$.

$$Q(j \mid x) = P(\phi(x, \mathcal{L}) = j) \qquad \textit{Equation 8:53}$$

$Q(j \mid x)$ is the relative frequency that the assigned class j for x is realized by ϕ. After consideration of the probability $P(j|x)$, the probability for a correctly classified class j at x is

$$\sum_j Q(j|x)\ P(j|x) \qquad \textit{Equation 8:54}$$

This probability needs to be maximized in terms of achieving significant results.

$$\sum_j Q(j|x)\ P(j|x) \leq \max_j P(j|x) \qquad \textit{Equation 8:55}$$

and

$$Q(j|x) = \begin{cases} 1\ if\ P(j|x) = \min_i P(i|x) \\ 0\ else \end{cases} \qquad \textit{Equation 8:56}$$

A so-called order-correct classifier ϕ is given by

$$argmax_j Q(j|x) \approx argmax_j P(j|x) \qquad \textit{Equation 8:57}$$

In the case where x is more often selected into a specific class j, then ϕ predicts the class j more often for x in comparison to other classes. This, however, does not mean that the accuracy is more precise. The probability for an aggregated predictor of correctly classified x is

$$\sum_j I(\ argmax_i Q(i|x) = j) P(j|x) \qquad \textit{Equation 8:58}$$

This results in the correct classification probability for ϕ_A,

$$r_A = \int_{x \in C} \max_j P(j|x)\, P_x(dx) + \int_{x \in C'} \left[\sum_j I(\phi_A(x) = j) P(j|x) P_X(x) \right] \qquad \textit{Equation 8:59}$$

C represents the set of all possible x and $P_x(dx)$ is the probability distribution x. Still, the accuracy can be low. If, however, the predictor is order correct for the majority of instances of x, then the aggregation process is capable of producing satisfying results.

The function used in the code also relies on decision *TREE* stumps.

8.1.1.8 RANDOM FOREST

Similar to the BAGGING approach, where decision *TREE*s are used for the classification problem, the *RANDOM FOREST* also relies on this method. Introduced by Breiman (2001) the approach adds more randomness to the process of *TREE* construction. In general, the nodes of the *TREE*s are split among all variables. In a *RANDOM FOREST* approach, these nodes are split based on the best of a subset of predictors, which are randomly chosen at each node [Liaw and Wiener (2002)]. Multiple *TREE*s are grown at the same time, and then the best predictor for each subset is selected by vote. So many decision *TREE*s $h_k(x)$ form the *RANDOM FOREST*.

Breiman (2001) defines the method as a classifier consisting of a collection of *TREE* structures $\{h(x, \Theta_k), k = 1, \dots\}$ where $\{\Theta_k\}$ are independent identically distributed random vectors and x is selected based on a unit vote from the classifiers for the most popular class. According to the author, the method seems counterintuitive, yet, it is able to outperform other methods such *SVM* or SLDA, and is further protected against overfitting. I have made a similar observation in this study (section 5.6). Other advantages are that *RANDOM FOREST* only needs a low number of parameters which are required for the construction of the classifier and that the method can easily handle high-dimensional data.

Following the formal definition by Breiman (2001) an ensemble of classifiers is given, $h_1(x), h_2(x), \dots, h_K(x)$, with a randomly selected training set based on the distribution of the random vector Y, X, and the margin function is given by

$$mg(X,Y) = \text{ave}_k\, I(\, h_k(X) = Y) - \max_{j \neq Y} \text{ave}_k\, I(\, h_k(X) = j) \qquad \textit{Equation 8:60}$$

$I(\cdot)$ is an indicator function for the margin, which estimates the average number of votes at X, Y. A large margin underlines the confidence in the assigned class. Frome here a generalization error is defined by

$$PE^* = P_{X,Y}(mg(X,Y) < 0) \qquad \textit{Equation 8:61}$$

with the probability $P_{X,Y}$ covering the whole space of X, Y. The Law of large Numbers states that with an increase in *TREE*s all sequences of $\Theta_k \dots PE^*$ will converge to

$$P_{X,Y}(P_{\Theta}(h(X,\Theta)=\mathrm{Y})-\max_{j\neq Y}P_{\Theta}(h(X,\Theta)=j)<0) \qquad \textit{Equation 8:62}$$

Equation 8:62 also illustrates that the *RANDOM FOREST* approach does not over fit when more *TREE*s are added.

The two essential measures for the *RANDOM FOREST* approach are the accuracy of the classifiers and identification of how independent they are (correlation). Using these for defining an upper bound for the classification, based on the generalization error and the margin function (Equation 8:60), the strength of each classifier is estimated by

$$s = E_{X,Y}mg(X,Y) \qquad \textit{Equation 8:63}$$

Considering Chebychev's inequality and assuming that $s \geq 0$,

$$PE^{*} \leq \frac{var(mg)}{s^{2}} \qquad \textit{Equation 8:64}$$

For the second parameter, the raw margin function is considered:

$$rmg(\Theta,\mathrm{X},\mathrm{Y}) = I(h(X,\Theta)=\mathrm{Y}) - I\big(h(X,\Theta)=\hat{j}(X,Y)\big) \qquad \textit{Equation 8:65}$$

A modified margin functions as

$$mg(X,Y) = E_{\Theta}\big[I(h(X,\Theta)=\mathrm{Y}) - I\big(h(X,\Theta)=\hat{j}(X,Y)\big)\big] \qquad \textit{Equation 8:66}$$

This can, therefore, be seen as the expectation of $rmg(\Theta,\mathrm{X},\mathrm{Y})$. If in an identity framework Θ and Θ' are independent with the same distribution, the margin function becomes

$$mg(X,Y)^2 = E_{\Theta,\Theta'} rmg(\Theta,\mathrm{X},\mathrm{Y}) rmg(\Theta',\mathrm{X},\mathrm{Y}) \quad \textit{Equation 8:67}$$

which results in

$$var(mg) = E_{\Theta,\Theta'}(p(\Theta,\Theta') sd(\Theta) sd(\Theta') \quad \textit{Equation 8:68}$$

with $p(\Theta,\Theta')$ the correlation and sd the standard deviation, between the two raw margin functions. Fixing Θ,Θ' with the correlation Θ with the standard deviation it can be concluded that

$$var(mg) \leq \bar{p} E_{\Theta} var(\Theta) \quad \textit{Equation 8:69}$$

with $\bar{p}$ the mean value of the correlation. Further, deriving

$$E_{\Theta} var(\Theta) \leq 1 - s^2 \quad \textit{Equation 8:70}$$

finally defines the upper bound for the generalization error as

$$PE^* \leq \frac{\bar{p}(1-s^2)}{s^2} \quad \textit{Equation 8:71}$$

The aim is to minimize Equation 8:71 for better results. The algorithm further applies the classification rule that the strength should be above 0.5 which is a similar approach to the weak learner boundary.

RANDOM FOREST approaches can also be modified with different kernel parameters, which will improve the overall performance of the classifier. However, it seems that the inbuilt

functions of the algorithm adjust on their own [Liaw and Wiener (2002)]. This is quite satisfying since it eases the handling.

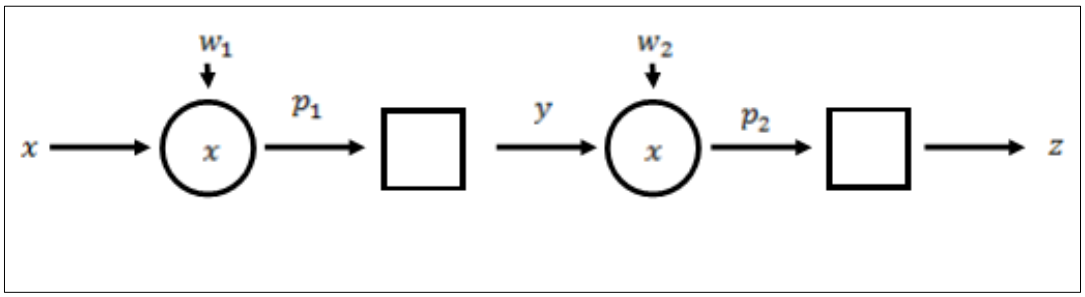
8.1.1.9 NEURAL NETWORKS (*NNET*)

Neural networks are seen by many experts as the most promising algorithm. Initially, the algorithm was influenced by biology and the neurons in the human brain. In the early 1940s with the beginning of computer calculations, researchers thought that a computer could be similar to the human brain or at least to its functioning.

Neurons are cells which are responsible for the information exchange and the interpretation of stimuli from our environment. Given its long-lasting background, this short explanation of the methodology only scratches the surface of the topic. Vast applications of neural networks have been performed in many fields, for example, picture recognition or music composition.

It is disappointing that the algorithm did not produce any satisfying results in this study. I assume that further adjustments to the code would have been necessary. The applied code relies on Venables and Ripley (2002), who present a formal definition of neural networks.

Figure 8:9 - Simple neural network consisting of two neurons



Note 8.39: The figure illustrates the functionality of a simple neural network. The above-presented scheme consists of two neurons, which try to modify the input x by applying weights w to it. The goal is it to generate a more or less similar output z by this modification.

The general idea is to train neural nets to create an outcome which is similar to the one desired. Following this, it can be stated that the input vectors $x_1, x_2, \dots, x_m$ enter a modification process which is dominated by some weights w_i and a threshold T_i, before an output z_i is produced (Figure 8:9).

$$\bar{z} = f(\bar{x}, \bar{w}, \bar{T}) \qquad \textit{Equation 8:72}$$

The illustrated process in Figure 8:9 can also be described as a function approximator. Equation 8:72 states a mathematical complex problem set, which can be simplified. A preferred way would be

$$\bar{d} = g(\bar{x}) \qquad \textit{Equation 8:73}$$

with $\bar{d}$ being the data. To estimate the difference between $\bar{z}$ and $\bar{d}$ the following performance function can be used to measure the magnitude of the difference:

$$P = -||\bar{d} * \bar{z}\,|| \qquad \textit{Equation 8:74}$$

The closer the value is to zero the better is the performance. Since weights and the threshold also influence the outcome of the classification or learning process, both need to be defined as well. One way of improving the performance is by representing the input parameters as partial derivatives:

$$\Delta\bar{w} = r\left(\frac{\partial P}{\partial w_1}i + \frac{\partial P}{\partial w_2}j\right) \qquad \textit{Equation 8:75}$$

The problem with Equation 8:75 is that a linear application to a non-linear space would not result in any acceptable results. It would be better to express $\bar{z}'$ as a function of $\bar{x}'$ and $\bar{w}'$. For this T the threshold will be set equal to w_0, with $w_0 = -1$, so that the reaction of the neuron can be measured right at the centre and the threshold disappears from the mathematical function. Further the smoothing parameter $\frac{1}{1+e^{-\alpha}}$ is introduced. If the basic concept is extended and the generated output of one of the neurons enters another neuron then Equation 8:74 can be rewritten. The simplest neural net is formed out of two neurons.

$$P = -\frac{1}{2}(d * z)^2$$ *Equation 8:76*

Now the chain rule for the partial derivatives can be applied. Here the individual steps in a simple neural network are derived. Figure 8:9 illustrates the individual steps.

$$\frac{\partial P}{\partial w_2} = \frac{\partial P}{\partial z} * \frac{\partial z}{\partial w_2}$$ *Equation 8:77*

which can be rewritten as

$$\frac{\partial P}{\partial w_2} = \frac{\partial P}{\partial z} * \frac{\partial z}{\partial p_2} * \frac{\partial p_2}{\partial w_2}$$ *Equation 8:78*

The whole process can be derived,

$$\frac{\partial P}{\partial w_1} = \frac{\partial P}{\partial z} * \frac{\partial z}{\partial p_2} * \frac{\partial p_2}{\partial y} * \frac{\partial y}{\partial p_1} * \frac{\partial p_1}{\partial w_1}$$ *Equation 8:79*

The partials of Equation 8:78 are defined as

$$\frac{\partial P}{\partial z} = d - z$$

$$\frac{\partial p_2}{\partial w_2} = y$$ *Equation 8:80*

where $\frac{\partial z}{\partial p_2}$ is a hidden function in the threshold box (the empty boxes in Figure 8:9).

$$\beta = \frac{1}{1 + e^{-\alpha}}$$

$$\frac{\partial \beta}{\partial \alpha} = \frac{d}{\partial \alpha}(1 - e^{-\alpha})^{-1} \qquad \textit{Equation 8:81}$$

$$\frac{\partial \beta}{\partial \alpha} = \beta(1 - \beta) = \frac{\partial z}{\partial p_2}$$

The above-described form is a feed-forward neural network. However, other forms have been developed, such as recurrent, recursive or deep belief neural networks with multiple cross-combinations among the individual neurons.

Medhat et al. (2013) briefly describe that the application of neural networks to text documents are based on the word frequency over the training dataset.

Table 8:31 - Robustness check I (all)

		MSCI Office City			*MSCI* office Mid-Town & West End		
		(1) *AFINN*_articles	(2) *BING*_Articles	(3) Maximum Entropy (1)	(1) *AFINN*_articles	(2) *BING*_Articles	(3) Maximum Entropy (1)
z_*AFINN*_article = L,	Standardized values for the lexicon approach with the *AFINN* lexicon	-0.731*** [0.143]			-0.633*** [0.132]		
z_*BING*_article = L,	Standardized values for the lexicon approach with the *BING* lexicon		-0.764*** [0.138]			-0.678*** [0.139]	
z_ceqart_max	Standardized values for the *MAXENT* algorithm based on the equalized training corpus with 3 categories			-0.664*** [0.137]			-0.691*** [0.150]
Constant		-0.908*** [0.135]	-0.958*** [0.142]	-0.866*** [0.131]	-1.104*** [0.145]	-1.179*** [0.154]	-1.115*** [0.147]
Observations		144	144	144	144	144	144
Log-likelihood		-59.69	-57.9	-63.35	-53.41	-51.11	-52.95
LR Chi2		33.18	36.75	28.31	26.12	27.55	27.03
Lag		2	2	0	0	1	0
pseudo-R-squared		0.218	0.241	0.183	0.196	0.212	0.203
AIC		123.371	138.823	130.708	110.816	106.211	109.896
BIC		129.311	144.763	136.647	116.755	112.151	115.836
Correctly classified (%)		34.38	76.39	78.47	84.03	82.64	84.03
Sensitivity		95.54	18.18	21.21	28	17.39	20
Specificity		81.94	93.69	95.5	95.8	95.04	97.48
Hosmer-Lemeshow χ^2		8.6	6.51	4.52	6.34	8.83	4.34
Prob > χ^2		0.376	0.590	0.807	0.609	0.357	0.822
area under Receiver Operating Characteristic (*ROC*) curve		0.816	0.771	0.801	0.817	0.835	0.808

Standard errors in brackets (*** p<0.01, ** p<0.05, * p<0.1)

Note 8.40: The table illustrates the probit results for the robustness check I for the full news corpus. Panel 1 uses the MSCI office city series as a dependent variable, while panel 2 uses the MSCI office Mid-Town & West End series. All three textual sentiment indicators are highly significant at a 1% level in both panels. The BING series, for the MSCI Mid-Town and West End probit model, generates the best results, according to the pseudo-R-squared value.

Table 8:32 - Robustness Check 1 (no housing)

		MSCI Office City			*MSCI* office Mid-Town & West End		
		(1) *AFINN*_articles	(2) *BING*_Articles	(3) Maximum Entropy (1)	(1) *AFINN*_articles	(2) *BING*_Articles	(3) Maximum Entropy (1)
z_*AFINN*_article = L,	Standardized values for the lexicon approach with the *AFINN* lexicon	-0.703*** [0.149]			-0.698*** [0.149]		
z_*BING*_article = L,	Standardized values for the lexicon approach with the *BING* lexicon		-0.900*** [0.169]			-1.301*** [0.248]	
z_ceqart_max = L,	Standardized values for the *MAXENT* algorithm based on the equalized training corpus with 3 categories			-0.311** [0.123]			-0.357*** [0.133]
Constant		-0.897*** [0.133]	-0.951*** [0.141]	-0.799*** [0.120]	-1.198*** [0.153]	-1.508*** [0.212]	-1.056*** [0.134]
Observations		144	144	144	144	144	144
Log-likelihood		-61.85	-55.54	-72.87	-49.73	-35.63	-59.39
LR Chi2		28.86	41.47	6.81	27.04	55.23	7.715
Lag		2	1	2	2	2	2
pseudo-R-squared		0.189	0.272	0.044	0.214	0.437	0.061
AIC		127.693	115.086	149.745	103.453	75.266	122.777
BIC		133.633	121.026	155.685	109.393	81.206	128.717
Correctly classified (%)		81.940	83.330	79.170	88.190	88.890	84.030
Sensitivity		31.250	40.630	6.250	30.430	47.830	0.000
Specificity		96.430	95.540	100.000	99.170	96.690	100.000
Hosmer-Lemeshow χ^2		10.660	15.640	12.370	5.090	3.680	0.982
Prob > χ^2		0.222	0.048	0.135	0.748	0.885	0.278
area under Receiver Operating Characteristic (*ROC*) curve		0.764	0.831	0.602	0.796	0.913	0.646

Standard errors in brackets (*** p<0.01, ** p<0.05, * p<0.1)

Note 8.41: The table illustrates the probit results for the robustness check I for the no housing sub-corpus. Panel 1 uses the MSCI office city series as a dependent variable, while panel 2 uses the MSCI office Mid-Town & West End series. The AFINN and BING indicators remain highly significant at a 1% level in both panels, while the MAXENT I model is significant at the 5% for the city series and highly significant for the Mid-Town & West End series. Again, the BING series, for the MSCI Mid-Town and West End probit model, generates the best results, according to the pseudo-R-squared value.

Table 8:33 - Robustness Check 1 (London)

		MSCI Office City			*MSCI* office Mid-Town & West End		
		(1) *AFINN*_articles	(2) *BING*_Articles	(3) Maximum Entropy (1)	(1) *AFINN*_articles	(2) *BING*_Articles	(3) Maximum Entropy (1)
z_*AFINN*_article = L,	Standardized values for the lexicon approach with the *AFINN* lexicon	-0.741*** [0.163]			-1.141*** [0.216]		
z_*BING*_article = L,	Standardized values for the lexicon approach with the *BING* lexicon		-0.815*** [0.164]			-1.051*** [0.190]	
z_ceqart_max	Standardized values for the *MAXENT* algorithm based on the equalized training corpus with 3 categories			-0.672*** [0.181]			-0.471*** [0.129]
Constant		-0.625*** [0.139]	-0.644*** [0.143]	-0.601*** [0.135]	-0.967*** [0.170]	-1.122*** [0.185]	-0.900*** [0.146]
Observations		111	111	111	111	111	111
Log-likelihood		-53.16	-50.32	-57.24	-36.03	-34.16	-49.78
LR Chi2		27.02	32.7	18.86	46.35	44.95	13.72
Lag		2	1	2	0	2	2
pseudo-R-squared		0.203	0.245	0.141	0.391	0.397	0.121
AIC		126.537	114.034	149.767	102.611	74.200	122.796
BIC		132.477	119.974	155.706	108.550	80.139	128.736
Correctly classified (%)		81.940	82.460	79.170	87.500	89.580	84.030
Sensitivity		31.250	40.630	6.250	40.000	52.170	0.000
Specificity		96.430	94.640	100.000	97.480	96.690	100.000
Hosmer-Lemeshow χ^2		11.450	16.490	12.380	7.110	3.870	9.830
Prob > χ^2		0.178	0.036	0.135	0.524	0.868	0.277
area under Receiver Operating Characteristic (*ROC*) curve		0.770	0.834	0.602	0.805	0.805	0.916

Standard errors in brackets (*** p<0.01, ** p<0.05, * p<0.1)

Note 8.42: The table illustrates the probit results for the robustness check I for the London sub-corpus. Panel 1 uses the MSCI office city series as a dependent variable, while panel 2 uses the MSCI office Mid-Town & West End series. All three textual sentiment indicators remain highly significant at a 1% level in both panels. Again, the BING series, for the MSCI Mid-Town and West End probit model, generates the best results, according to the pseudo-R-squared value.

Table 8:34 - Robustness Check 1 (100,000)

		MSCI Office City			*MSCI* office Mid-Town & West End		
		(1) *AFINN*_articles	(2) *BING*_Articles	(3) Maximum Entropy (1)	(1) *AFINN*_articles	(2) *BING*_Articles	(3) Maximum Entropy (1)
z_*AFINN*_article = L,	Standardized values for the lexicon approach with the *AFINN* lexicon	-0.706*** [0.134]			-0.855*** [0.159]		
z_*BING*_article = L,	Standardized values for the lexicon approach with the *BING* lexicon		-1.053*** [0.173]			-1.237*** [0.205]	
z_ceqart_max	Standardized values for the *MAXENT* algorithm based on the equalized training corpus with 3 categories			-0.810*** [0.148]			-0.977*** [0.176]
Constant		-0.878*** [0.133]	-0.983*** [0.149]	-0.918*** [0.139]	-1.175*** [0.155]	-1.405*** [0.195]	-1.257*** [0.170]
Observations		144	144	144	144	144	144
Log-likelihood		-60.940	-49.390	-58.410	-46.47	-34.72	-43.84
LR Chi2		33.150	56.240	38.190	39.99	63.49	45.26
Lag		0	0	0	0	0	0
pseudo-R-squared		0.214	0.363	0.246	0.301	0.478	0.340
AIC		125.875	102.781	120.830	96.937	73.441	91.672
BIC		131.814	108.721	126.769	102.876	79.380	97.611
Correctly classified (%)		81.250	85.420	80.560	89.580	89.580	86.810
Sensitivity		30.300	54.550	33.330	44.000	64.000	44.000
Specificity		96.400	94.590	94.590	99.160	94.960	95.800
Hosmer-Lemeshow χ^2		12.940	10.750	17.190	7.800	10.910	12.100
Prob > χ^2		0.114	0.228	0.028	0.454	0.207	0.147
area under Receiver Operating Characteristic (*ROC*) curve		0.830	0.881	0.855	0.849	0.916	0.895

Standard errors in brackets (*** p<0.01, ** p<0.05, * p<0.1)

Note 8.43: The table illustrates the probit results for the robustness check I for the 100,000 sub-corpus. Panel 1 uses the MSCI office city series as a dependent variable, while panel 2 uses the MSCI office Mid-Town & West End series. All three textual sentiment indicators remain highly significant at a 1% level in both panels. Again, the BING series, for the MSCI Mid-Town and West End probit model, generates the best results, according to the pseudo-R-squared value.

Table 8:35 - Robustness Check 1 (FT)

		MSCI Office City			*MSCI* office Mid-Town & West End		
		(1) *AFINN*_articles	(2) *BING*_Articles	(3) Maximum Entropy (1)	(1) *AFINN*_articles	(2) *BING*_Articles	(3) Maximum Entropy (1)
z_*AFINN*_article = L,	Standardized values for the lexicon approach with the *AFINN* lexicon	-0.576*** [0.136]			-0.607*** [0.144]		
z_*BING*_article = L,	Standardized values for the lexicon approach with the *BING* lexicon		-0.697*** [0.151]			-0.827*** [0.173]	
z_ceqart_max	Standardized values for the *MAXENT* algorithm based on the equalized training corpus with 3 categories			-0.204* [0.118]			-0.171 [0.120]
Constant		-0.865*** [0.129]	-0.920*** [0.136]	-0.777*** [0.118]	-1.163*** [0.149]	-1.271*** [0.166]	-1.011*** [0.128]
Observations		144	144	144	144	144	144
Log-likelihood		-65.71	-62.83	-74.67	-53.02	-47.82	-62.26
LR Chi2		21.13	26.9	3.207	20.45	30.85	1.966
Lag		2	1	2	2	2	2
pseudo-R-squared		0.138	0.176	0.021	0.162	0.244	0.015
AIC		135.429	129.659	153.349	110.043	99.641	128.526
BIC		141.368	135.599	159.288	115.983	105.580	134.465
Correctly classified (%)		79.170	81.940	78.470	84.720	86.110	83.330
Sensitivity		15.630	31.250	3.130	8.700	26.090	0.000
Specificity		97.320	96.430	100.000	99.170	97.520	99.170
Hosmer-Lemeshow χ^2		10.900	7.410	7.790	7.990	18.120	9.910
Prob > χ^2		0.208	0.493	0.455	0.435	0.020	0.272
area under Receiver Operating Characteristic (*ROC*) curve		0.755	0.770	0.587	0.800	0.823	0.630

Standard errors in brackets (*** p<0.01, ** p<0.05, * p<0.1)

Note 8.44: The table illustrates the probit results for the robustness check I for the Financial Times sub-corpus. Panel 1 uses the MSCI office city series as a dependent variable, while panel 2 uses the MSCI office Mid-Town & West End series. Both the AFINN and the BING series remain highly significant at a 1% level, while the MAXENT I sentiment measure is only significant at a 10% level in the first panel. Again, the BING series, for the MSCI Mid-Town and West End probit model, generates the best results, according to the pseudo-R-squared value.