Information Processing and Stock Market Volatility – Evidence from Real Estate Investment Trusts

Completed Research Paper

Jan-Otto Jandl
University of Freiburg
jan-otto.jandl@vwl.uni-freiburg.de

Abstract

This study proposes novel measurements of investor psychology utilizing Big Data and reveals the impact of different measures on asset price volatility. We construct a news sentiment measure based on news articles reflecting information supply. News content is transformed into quantitative data utilizing sentiment analysis. We further investigate a metric for investor attention based on search queries from the web representing information demand. Consequently, we investigate how asset price volatility is attributable to information processing by investors. The main contribution of this paper is the analysis of novel insights into information processing by investors in the presence of Big Data. In particular, we find that the impact of news content (information supply) generally is much stronger than the effect of Big Data search behavior (information demand). While our results confirm the negativity-bias of investors, we find that measures additionally incorporating positive content outperform measures solely based on negative connotations.

Keywords

Big Data Analytics, Text Mining, Economics of Information Systems.

Introduction

It has long been posited that investor psychology constitutes a crucial determinant for the price formation in asset markets (Keynes 1936). In particular, speculative stock price bubbles seem to be attributable to overconfidence among traders (DeBondt and Thaler 1985) and psychological biases between market participants (Hirshleifer 2001). Although there has been much discussion and interest in the role of mass psychology in financial markets, empirical evidence has long been limited due to the lack of appropriate measures for investor sentiment (Akerlof and Shiller 2009).

Within the last decade, Big Data has opened the path for research incorporating extensive volume, variety and velocity of information. While volume and variety refer to the immense amounts and the heterogeneity of data published in various sources, velocity reflects the speed of data arrival in the present information age. An increasing branch of Information Systems research has utilized these novel dimensions of information supply. For instance, research on the relations between media content and corresponding market reactions has provided new insights into information processing in financial markets. Prior work adopting sentiment analysis suggests an asymmetric impact of investor sentiment on asset prices. That is, the effect of media sentiment is particularly pronounced incorporating negative connotations (Tetlock 2007) and additionally much stronger in times of recessions (García 2013).

In addition to information supplied by the media, we suppose that the attention of individuals to market-specific information, hence information demand, similarly instructs trading behavior and stock price volatility. Nevertheless, most proxies for investor attention are infrequent and not duly in reporting investor psychology. As user generated content (UGC) arguably reflects trading motivation of investors, active

search behavior unravels the attention of individuals to specific topics and provides an additional eligible proxy for investor sentiment (Da et al. 2011). In particular search data from the web provides an adequate measure for aggregate attention to specific topics and has been proven as a valuable tool in economic research (Choi and Varian 2012).

We establish novel measures of investor psychology as innovative metrics to analyze information processing and decision making in stock markets. We incorporate both factors, media content as well as search query data as potential proxies for investor sentiment and therefore investigate the sensitivity and attention of investors to information on a specific market simultaneously. We construct a news sentiment measure based on articles from the Thomson Reuters News Archive representing information supply. News content is converted into quantitative data utilizing sentiment analysis. We further investigate sentiment metrics based on search queries from the web reflecting information demand in order to identify the attention of individuals to stock related content.

This study is motivated by the increasing empirical evidence for the role of psychological bias in financial markets. While the majority of prior work focuses on the general stock market, empirical evidence for listed real estate stocks has been very limited. Although prior literature stresses the relevance of investor sentiment in securitized real estate markets, the proxies are mostly not timely and infrequent and therefore seem to reflect psychological bias of investors just indirectly. Compared to these proxies, we propose that sentiment metrics on the basis of information variables are possibly better approaches to derive sentiment.

We select the securitized real estate market for a number of reasons. First, the real estate sector constitutes a major asset of individual households and therefore plays a significant role in the global economy. As the unexpected decline in property prices has widely been believed to be a precursor of the subsequent Great Recession (Stiglitz 2009), we consider insights into information processing in real estate stock markets to be of great importance for the global economy. Secondly, the securitized real estate market constitutes an asset closely associated with either, the stock market and the underlying property market.

Therefore, this study aims at providing an additional cognition about the role of information with potential feedback to stock markets and real assets utilizing Big Data. Consequently, we investigate how stock returns are attributable to several proxies for investor psychology. In order to remove white noise, we particularly focus on news and web search data on the real estate market and investigate the return volatility of Real Estate Investment Trusts (REITs). Therefore, we are able to identify and analyze information variables and corresponding market reactions. We examine the links between different measures for information affecting investor psychology and stock markets and explicitly investigate the influence of information processing on decision making in financial markets.

The remainder of this paper is structured as follows. Section 2 provides a summary of the related literature on behavioral economics, sentiment analysis and real estate stocks. Section 3 describes the data sources and data mining techniques applied in this study. Section 4 presents the empirical analysis and evaluates the quantitative results. Section 5 contains critical remarks and concludes.

Literature Review

Though the concept of expected utility for investigating decisions under risk has widely been accepted as a normative model of rational behavior, more recent literature stresses that psychological aspects are of tremendous importance for investment decisions under uncertainty (Kahneman and Tversky 1979). Hence, literature on psychology and economics suggests that investor psychological bias or 'animal spirits' unravel irrational behavior of individual investors.

The empirical evidence and contributions of the increasing research corpus addressing this notion prove that sentiment analysis has become an effective and reliable tool in economics. The majority of recent financial text mining research performs word analysis based on predetermined dictionaries (Demers and Vega 2010; Henry 2008; Jegadeesh and Di Wu 2013; Tetlock 2007). These dictionaries originate from research in both the psychology and finance discipline and are used to count the frequency of pre-defined positive and negative words. This approach is straightforward and produces reliable results. As sentiment analysis is applied to a wide range of domains and text sources, research has devised various approaches to measure sentiment (Pang and Lee 2008). As machine learning approaches may suffer from overfitting (Antweiler and Frank 2004), we focus on dictionary-based sentiment extraction instead.

However, the sole investigation of news content does not properly account for the consecutive attention of investors. This is even more important in the presence of Big Data in which the variety and intensity of information induces a lack of attention. Therefore, prior work proposes additional measure to quantify the interest of investors in specific industries. Prominently Google Trends data is found to significantly impact volatility and trading activity in financial markets. Da et al. (2014) propose a novel and more direct metric of investor attention based on the frequency of Google queries. They find that the search volume for stock tickers serves as an appropriate indicator for investor attention to stocks and can be used as a predictor for returns, in particular for initial public offerings (IPOs). Accordingly, they stress that measures based on search volume are much more timely in capturing the attention of investors and retail investors.

Recent work additionally highlights the predictive value of search query data. Preis et al. (2013) stress that massive new data sources may offer new insights into the current state as well as future trends of the economy. By estimating search patterns, they illustrate that query data exhibits warning signs of stock price variations. Choi and Varian (2012) show that incorporating the Search Volume Index (SVI) helps improving the prediction of several economic measures. Drake et al. (2012) find that search volume generally starts increasing around two weeks prior to earnings announcements and preempts information content of the subsequent announcement. While the prior studies use weekly data, Da et al. (2014) show that daily sentiment-related queries are an effective measure for investor sentiment and predict market movements.

This study complements the existing studies in several ways. While the majority of research focuses on financial markets in the United States, we provide a comparative analysis additionally investigating the United Kingdom. Furthermore, we incorporate measures for both information supply and information demand. The rationale is provided by Da et al. (2011), who state that investor attention as represented by the SVI does not seem to exhibit notable correlations with news-based sentiment measures.

Data

We gather data from several sources. We first describe stock market data including real estate stocks and stock price indices. Second, we present news sentiment data, which is extracted using text mining from machine-readable news articles. Third, we present the derivation of an aggregate data source for individual attention to information from Google search data and report the corresponding sample statistics.

Financial Market

Stock market data for UK and US REITs is based on the FTSE EPRA/NAREIT Global Real Estate Index Series data provided by the European Public Real Estate Association (EPRA) in collaboration with the Financial Times Stock Exchange (FTSE) and the National Association of Real Estate Investment Trusts (NAREIT). We obtain the total return indices as a measure of investment performance. We let R_t^c denote the log return of the REIT index of the respective country *c*. Information on business cycles is gathered from the National Bureau of Economic Research (NBER). We define a dummy variable D_t which takes the value of one if and only if date *t* is considered to be within recession period as defined by the NBER.

Table 1 reports the unconditional sample statistics for the daily log-returns of the Total Return Indices. The statistics are provided for the entire sample period as well as NBER expansion and recession periods. Panel A indicates that the UK REIT return for the entire sample period with 1.2 basis points was marginally positive, with a daily volatility of 190 basis points. The sample return changes across NBER expansions and recessions, which comprise 2,024 respectively 406 days out of the 2,430 trading days considered in our sample. While the daily return was 6.6 basis points during expansions, it was -25.6 basis points in times of economic recession. Similar patterns are observed for US REIT returns as punctuated by the sample statistics provided in Panel B. The average daily return over the entire sample period was 4.3 basis points which was accompanied by a daily volatility of 239 basis points. The return was 8.6 basis points during expansions and -17.4 basis points during recessions. The entire sample for the US REIT returns consists of 2,374 trading days of which 397 respectively 1,977 days have been considered as recessions and expansions by the NBER. The sample statistics additionally highlight the changing daily volatility over the business cycle. While the standard deviation is observed to be relatively low during expansions, it is more than twice that during recessions in both countries. We note that the marked disparities in observations for the two markets are due to public holidays which partly differ in the UK and the US.

Dates	Mean	Std. dev.	25% -quan.	Median	75% -quan.	Mean	Std. dev.	25% -quan.	Median	75% -quan.
	Panel A: UK REIT Returns (R_t^{UK})					Panel B: US REIT Returns (R_t^{US})				
All	0.012	1.909	-0.828	0.069	0.898	0.043	2.396	-0.784	0.112	0.931
Exp.	0.066	1.443	-0.676	0.100	0.806	0.086	1.413	-0.643	0.139	0.870
Rec.	-0.256	3.373	-2.231	-0.178	1.764	-0.174	4.936	-2.304	-0.166	2.139

Table 1. Descriptive Statistics for REIT returns in the UK and the US

We specify a simple model of stock returns to uncover time-series characteristics, which account for the time variation in volatility. In essence, we estimate a model analogous to García (2013) in the form of

$$R_t^c = (1 - D_t)\beta_1^c L_S(R_t^c) + D_t\beta_2^c L_S(R_t^c) + \eta^c X_t + \varepsilon_t,$$

which is to be expanded in the empirical analysis. In this setting, D_t denotes a dummy variable indicating a NBER recession period, L_S the lag-operator of length s, X_t a set of independent variables and ε_t the zero-mean error term. We include a day-of-the-week and business cycle dummies as well as a constant term in X_t.

The regression results of the parsimonious model for a maximum lag length of s=5 with heteroscedasticity-consistent standard errors according to (White 1980) indicate no statistically significant autocorrelation for REIT returns in the UK. Contrary, there is strong evidence for negative autocorrelation for REIT returns in the US which may indicate some form of mean reversion. Corresponding to the results of García (2013), autocorrelation is particularly significant during expansions. The results additionally highlight significantly lower returns during recessions in both countries. While UK REIT returns on Monday are roughly between 26 and 29 basis points lower than on most other days, no significant day of the week effects are identified for the US.

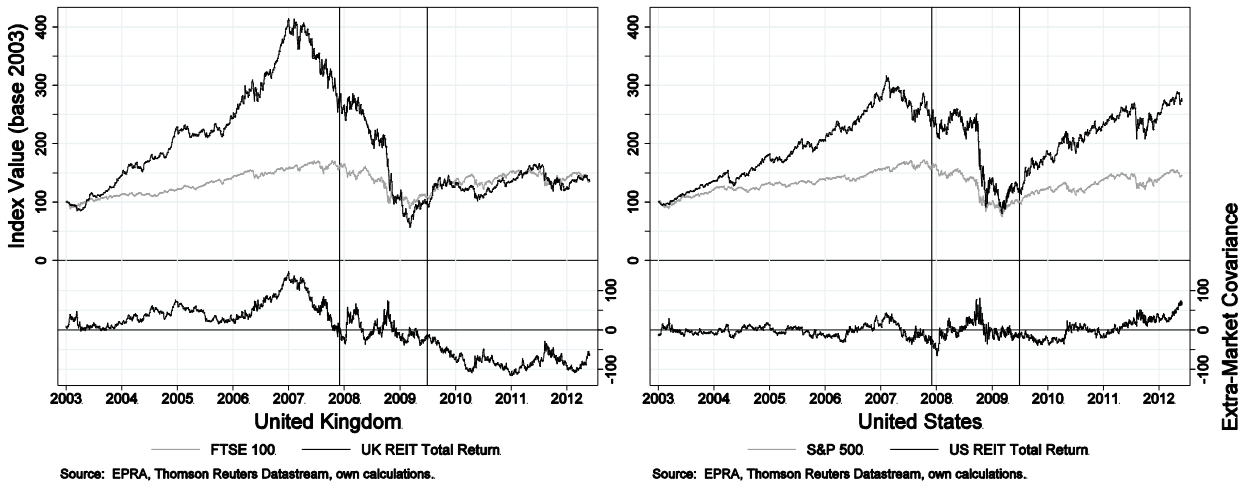

Figure 1. Stock Market and REIT Indices and Extra-Market Covariance

Prior work on real estate investments has indicated that traded real estate securities are more similar to other types of listed stocks than to the direct property market in the short run (Morawski et al. 2008). Therefore, it seems arguable if general stock market effects overstate the variation of REIT returns. Consequently, we calculate Pearson product-moment pairwise correlations between the REIT and stock price indices in order to identify if statistically significant short-run relations exist in our sample data. The coefficient measures the strength of linear dependence between two variables and is valued on a scale between -1 and 1. We use the S&P 500 index and the FTSE 100 index as stock price indices in the respective

countries. The pairwise correlations between the stock price and REIT indices are 0.91 in the US and 0.70 in the UK, statistically significant on the 0.001 level and hint at strong linear relations between securitized real estate investments and the general stock market in both countries. As the observed correlations are even stronger than in prior studies (McCue and Kling 1994), linear orthogonalisation is assumed to effectively remove stock market effects from REIT returns.

Analogous to (McCue and Kling 1994), we regress REIT returns on the general stock market index returns and use the residuals as an indicator for extra-market covariance. Potential issues arising through single REIT effects should be negligible due to diversification inherent in the weighted averages represented by the indices. The extra-market covariance of the residuals represents the pure real estate industry effect (Brooks and Tsolacos 2000). Figure 1 illustrates the derivation of the housing-industry effect based on the orthogonalisation of REIT returns. We observe that the time series of the extra-market covariance exhibits some similarities with the corresponding REIT index for the largest part of the sample period. However, we identify considerable differences within the recession period, which is illustrated through the area between the vertical reference lines. We conclude that the time series reflect two distinct measures for real estate-related returns which are to be included in the further econometric analysis. Further, the analysis indicates that the time-varying volatility of returns is still present after removing general stock market effects and that the residual real estate returns exhibit characteristics similar to REIT returns.

News Content

We aggregate media content measures as proxies for investor sentiment by applying sentiment analysis. Sentiment analysis generally includes a text corpus, which is preprocessed by linguistic tools and transformed into sentiment scores using various approaches. Hence, sentiment analysis represents an instrument, which converts qualitative information inherent in messages into quantifiable measures. The common architecture of a generic sentiment analysis system can be classified into corpus retrieval, document processing and the document analysis module. Due to increasing data accessibility, the retrieval of the text corpus can originate from a huge variety of sources such as newspaper articles, press releases, 10-K reports or diverse information published in social media.

We retrieve our news corpus from announcements published in the Thompson Reuters News Archive for Machine Readable News. This news corpus contains articles, which exclusively address novel third-party content. The announcements are filtered for real estate relevance using Reuters topic codes for real estate content (REA) in the United States (US) and Great Britain (GB). We exclude announcements that mirror the opinion of analysts or contain less than 50 words to avoid white noise. We identify a total of 18,629 announcements solely addressing the UK real estate market with a total of 5,932,525 words. Each day around seven announcements are published with an average of 320 words. Announcements focusing on the housing market in the US amount to 59,264 and contain 14,166,577 words on the whole. Around 22 announcements with an average word count of 243 are published each day.

Each announcement is subject to document processing which converts running texts into machine-readable tokens. The preprocessing phase consists of several operations which transfer the machine-readable news into a term-by-document matrix (Manning and Schütze 1999). The document processing phase includes the application of several linguistic tools in order to transform the corpus into features that are applicable to further quantitative analysis. A parsimonious procedure involves steps such as tokenization, stop word removal, synonym merging and stemming. *Tokenization* involves the decomposition of the text into elements, which are subsequently used as input for further analysis. Most commonly, texts are segmented into sentences or single words. Grefenstette and Tapanainen address the general problems involved in tokenization and discuss the highly subjective definition of tokens. In a second step, *stop word removal* excludes redundant words from the further analysis (Manning and Schütze 1999). The third step applies *synonym merging* and merges words with similar meanings into groups. The fourth step consists of *stemming* and reduces words to their stem in order to consolidate words containing similar meanings. A first stemming algorithm based on a dictionary of 294 commonly used suffixes has been introduced by Lovins (1968). As this algorithm contains several decoding and recursion rules, a simple and fast technique involving six steps is described in Porter (1980), which is still predominantly used in text analysis and therefore applied in this study. The results of the document processing phase are summarized in a term-by-document matrix as described by Manning and Schütze (1999).

The document analysis module attaches annotations to specified snippets of texts by applying a lexicon. The most crucial resource for sentiment analysis is the acquisition of the lexicon. While the manual approach leads to a lexicon coded by hand and therefore is highly subjective, the dictionary-based approach utilizes exogenous codes provided by predefined dictionaries. Commonly used dictionaries include the General Inquirer (GI) (Stone 1968), the Sentiment Lexicon (SL) (Hu and Liu 2004), the MPQU Subjectivity Lexicon (MP) (Wilson et al. 2005), the SentiWordNet (Esuli and Sebastiani 2006) and the Emoticon Lexicon (EL) (Mohammad and Turney 2010). Although dictionary-based approaches provide a more objective method, the latter dictionaries usually do not cater for the peculiarities of specific domains. Therefore, domain-specific lexica have been created endogenously through corpus-based approaches. Generally, the compilation of a domain-specific lexicon bases on the analysis of a large corpus of domain-specific documents by algorithms which typically parse the sentences and identify the associated sentiment expressions. We use the General Inquirer (GI) (Stone 1968) as a general psychological dictionary as well as the Loughran McDonald (LM) (Loughran and McDonald 2011) and Henry's Finance Dictionary (HE) (Henry 2008) as domain-specific lexica containing finance-specific words which have been extensively used in the literature (García 2013; Henry 2008; Henry and Leone 2009; Loughran and McDonald 2011, 2013; Tetlock 2007). Concluding, the dictionary-based approach represents a simple and the most intuitive application of text mining, which seems most applicable and is predominantly used in recent financial text mining research.

We let w_{it} denote the total amount of words in column i on date t and additionally count the positive p_{it}^d as well as negative words n_{it}^d according to the respective dictionary d applied. Based on these variables, daily measures of positive and negative media content reflect the portion of positive and negative words in the news. We define two distinct measures of media content, that is the fraction of negative words, referred to as **Negativity** (*Neg*) (Tetlock 2007) and calculated as

$$Neg_t^d = \frac{\sum_{it} n_{it}^d}{\sum_{it} w_{it}^d},$$

as well as Net **Sentiment** (*Sent*) or Net-Optimism (Henry and Leone 2009) in each days aggregate news announcements defined as

$$Sent_t^d = \frac{\sum_{it} p_{it}^d - \sum_{it} n_{it}^d}{\sum_{it} w_{it}^d}.$$

In essence, our news sentiment measures are specified as the fraction of negative words (*Neg*) and the difference between the portion of positive and negative words (*Sent*) in all qualifying news announcements. The descriptive statistics for the sentiment measures are summarized in Table 2. Note that all numbers are given in percentages.

	United Kingdom					United States				
Media Measure	Mean	Std. dev.	25% -quan.	Median	75% -quan.	Mean	Std. dev.	25% -quan.	Median	75% -quan.
Neg. (GI)	5.13	1.63	4.08	5.03	6.16	5.56	1.29	4.85	5.56	6.35
Neg. (HE)	0.78	0.55	0.42	0.71	1.06	0.91	0.50	0.55	0.86	1.21
Neg. (LM)	2.11	1.10	1.33	2.04	2.76	2.52	0.92	1.99	2.45	2.99
Sent. (GI)	3.78	2.68	2.34	3.94	5.38	4.33	1.88	3.28	4.36	5.42
Sent. (HE)	0.97	0.92	0.40	0.87	1.43	0.94	0.76	0.49	0.86	1.33
Sent. (LM)	-0.89	1.33	-1.63	-0.86	-0.07	-1.23	1.10	-1.86	-1.17	-0.57

Table 2. Descriptive Statistics for Sentiment Measures along the Business Cycle

The sample statistics covering 3,052 days uncover similar sentiment patterns for the UK and the US. The average number of negative words in daily news amounts to roughly 5% according to the GI with a daily variation of around 1.5 %. Finance-specific dictionaries identify much less negative words with a daily average around 2% (LM) and 1% (HE). This seems reasonable as the GI reflects a general psychological sentiment measure while the remaining have a much more narrow finance-specific context. The net sentiment measures indicate that while the general psychological news sentiment (GI) is clearly positive with around 4% more positive than negative words, finance-specific news sentiment with around 1% (HE) and -1% (LM) seems slightly balanced over the sample period. However, the HE and LM news measures highlight significant differences between the dictionaries. While the HE identifies just around half as much negative than positive words, the LM measure is almost twice as high regarding the net sentiment measure.

The sample statistics additionally highlight that the application of finance-specific dictionaries reduces the number of relevant words. In particular, the differences between the general psychological and finance-specific dictionaries will be of central interest as a large body of prior work in psychology stresses a negativity-bias. Accordingly, individuals are much more prone to negative than to positive information (Baumeister et al. 2001).

Preliminary tests indicate a reasonable pattern of correlation among the media content measures and in particular two insights, which are to be taken into account in the further analysis. That is, sentiment measures based on the GI and LM exhibit high correlations in terms of sign, magnitude and significance as is the case for net sentiment metrics and their respective negativity measures. We assume that the identification of dictionaries and metrics is fruitful for information processing in financial markets as in particular the finance-specific dictionaries reflect distinct content.

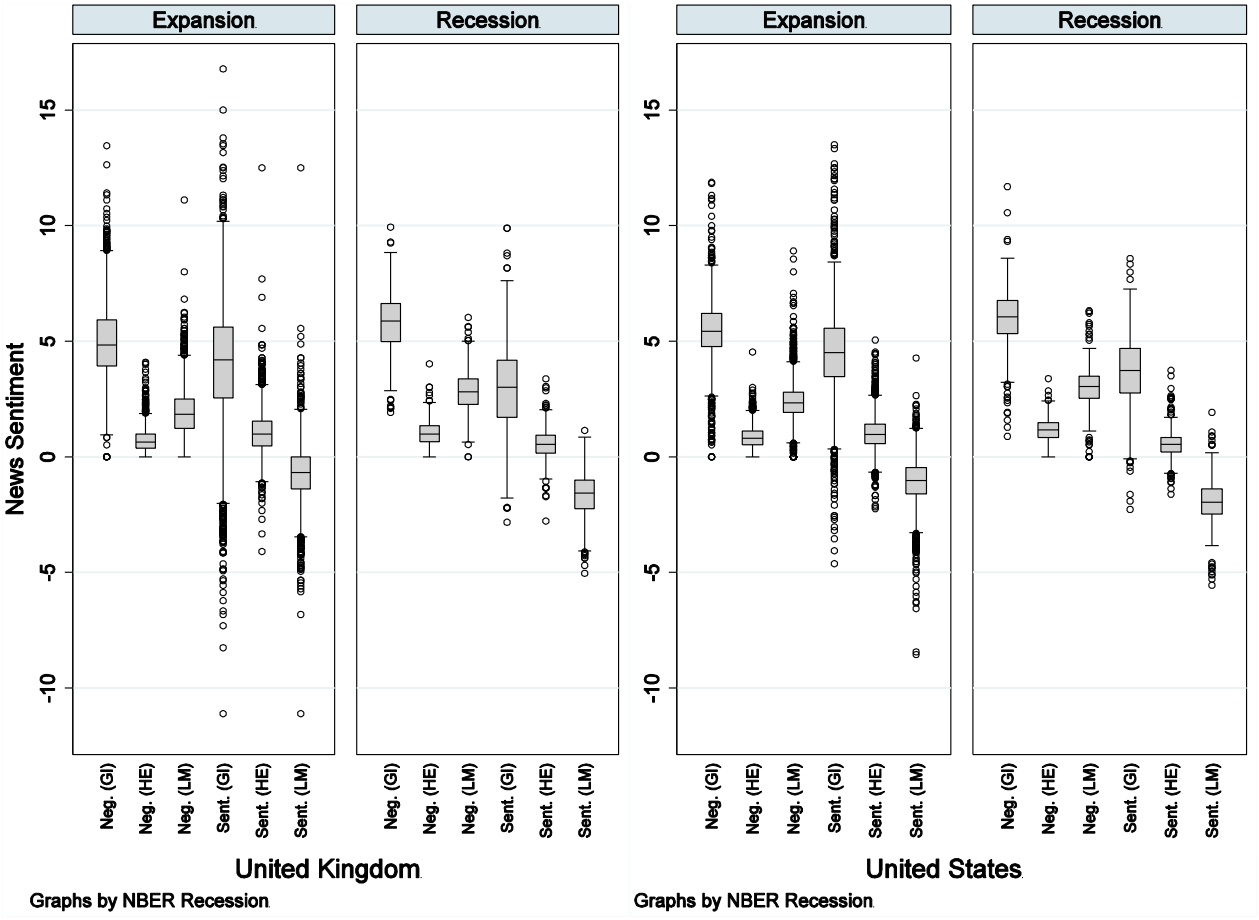

Figure 2. Box Plots of News Sentiment Measures

Figure 2 presents box plots for each of the metrics, Negativity (*Neg*) and Net-Sentiment (*Sent*), applied as a function of the NBER business cycle. We stress that the variation in news sentiment along the business cycle are clearly visible with *Neg* being lower and *Sent* being higher during recessions. As the volatility of all news sentiment measures is significantly different along the business cycle, we propose that volatility clustering has to be addressed in the econometric analysis.

Search Queries

Search query data is derived from Google Trends and represents millions of individuals concerned with real estate markets. We regard the collective 'swarm intelligence of Internet users' (Preis et al. 2013) as an important additional determinant for stock market returns. Google provides Search Volume Indices (SVI) of search terms and categories, which represent the number of searches in relation by its time-series average. The SVI is additionally available for specific geographical regions. We capture the attention of individuals to real estate markets in both countries separately based on two distinct measures. First, we examine the SVI for the category 'Property'. Google provides a classification of 25 top-level categories, which are determined using automated classification. Second, we estimate a list of search terms that potentially reveal the attention of individuals to information on the real estate sector. To do so, we use the list of top searches that Google provides for the search term 'Real Estate'. As these terms are closely related, we manually derive the SVI for each term in each country and construct a measure of revealed public interest in real estate markets.

However, the use of Google SVI data is accompanied by several caveats. First, the information on actual counts of search queries is held private. Instead, Google provides approximations of the actual search behavior by reporting the propensity of users to search for particular terms. In particular, values relative to the total number of searches in the corresponding time period ranging from 0 to 100 are reported. Hence, the data provided represents a sample of actual terms and may contain measurement errors. Second, the search data does not reveal the type of the individual investor who requests the information additionally adding noise to the sample data. Nevertheless, prior studies have addressed this issue and suggest that the majority of web searches are attributable to retail investors (Da et al. 2011). Third, while weekly data is available for the entire sample period covering the years between 2004 and 2013, daily data is only available for time intervals with a maximum time range of three months. In order to derive a time series data corresponding to the development of public attention to real estate-related topics over the entire sample period, we manually download the sample data for each quarter and linearly transform the time series data. We follow the procedure of Da et al. (2014) and scale the daily Google Trends data for each quarter by the maximum SVI in the respective quarter. We define the daily change in search volume for the search term or category j as

$$\Delta SVI_{j,t} = \ln\left(SVI_{j,t}\right) - \ln\left(SVI_{j,t-1}\right).$$

The corresponding time series of daily changes in SVI for the period between October 2007 and April 2008 are depicted in Figure 3. We note adjustments in web search behavior around the business cycle change in December 2007 and also observe a similar pattern for other dates for business cycle change. Apparently, considerable differences in the variation of SVI across terms and categories are present in each country which appear much in advance of the official declaration of a business cycle change by the NBER (December 2008). We additionally identify a weekly seasonality of the SVI in each country which is conforming the results of Da et al. (2014). In particular, we observe significant increases in SVI change on Monday and Tuesday and subsequent reversing effects.

In order to address heteroscedasticity and seasonality issues, we identify a normal level of attention and estimate the Abnormal change in SVI (ASVI) as described in Da et al. (2014). In particular, we remove seasonality by regressing $\Delta SVI_{j,t}$ on weekday and month dummies and scale each residual time series applying standardisation. We additionally winsorize the data at the 5 % level to mitigate issues arising from outliers. In essence, the normal level of search volume is removed to uncover abnormal search behavior. Hence, $\Delta ASVI_t$ represents a standardized, windsorized and deseasonalized change in daily search queries.

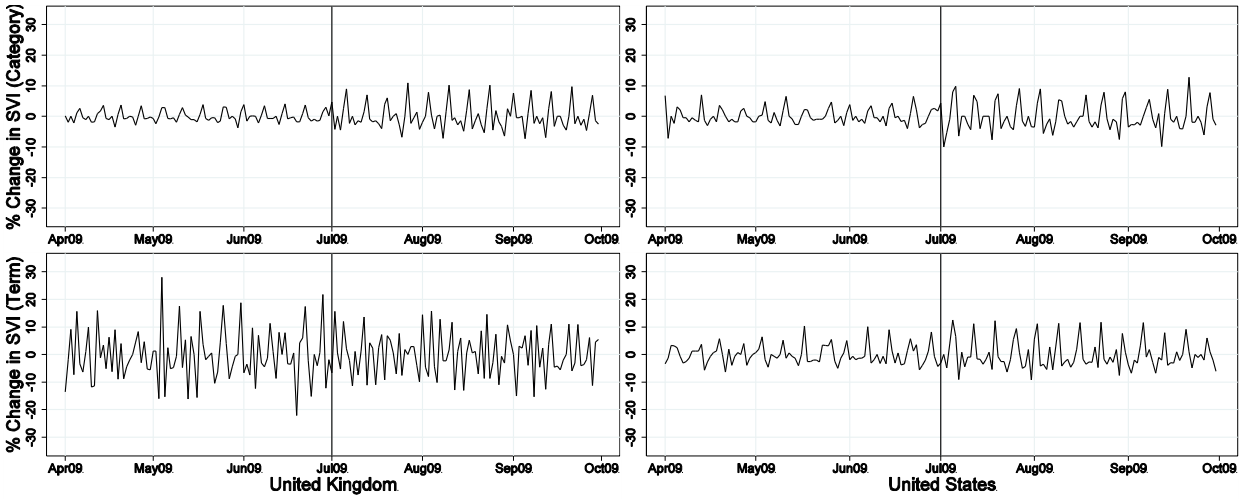

Figure 3. Daily Change in SVI for Real Estate Search Terms and Property Category

We calculate Pearson correlation coefficients in order to identify potential collinearity between the public attention variables $\Delta ASVI_t$. The correlations coefficients are reported in Tabe 3 and indicate strong relations in terms of significance and sign. That is, the invariably positive correlation coefficients suggest that the measures for changes in abnormal search behavior tend to move in a synchronous manner. Nevertheless, the magnitude of correlation is rather small. In fact, the correlation between the information demand measures in the UK is only marginally different from zero and just significant on the 10%-level.

Nevertheless, the corresponding correlation between the measures in the US amounts to roughly 0.46 and is statistically significant. We additionally report cross-country correlations for the search behavior. While the majority of correlations between the measures across countries are not significant in magnitude, we note a highly significant and notable positive correlation between changes in the ASVI measures based on the predefined 'Property' category.

	Web Search Variables				Media Content Variables					
	Terms (UK)	Categ. (UK)	Terms (US)	Categ. (US)	Neg. (GI)	Neg. (HE)	Neg. (LM)	Sent. (GI)	Sent. (HE)	Sent. (LM)
Terms	–				0.024	0.030	0.041	-0.016	-0.001	-0.010
(UK)					(0.267)	(0.166)	(0.059)	(0.464)	(0.971)	(0.653)
Category	0.034	–			-0.020	-0.041	-0.021	0.018	0.014	-0.002
(UK)	(0.059)				(0.366)	(0.058)	(0.336)	(0.411)	(0.519)	(0.934)
Terms	0.170	0.106	–		0.049	0.013	0.031	-0.012	-0.017	-0.020
(US)	(0.000)	(0.000)			(0.027)	(0.568)	(0.162)	(0.583)	(0.452)	(0.364)
Category	-0.083	0.479	0.459	–	0.032	0.017	0.052	-0.020	-0.042	-0.063
(US)	(0.000)	(0.000)	(0.000)		(0.141)	(0.443)	(0.019)	(0.360)	(0.056)	(0.004)

Table 3. Correlation among Measures for Information Supply and Information Demand

We additionally report the correlations between the search variables and the corresponding media content variables in each country. Interestingly, we do not detect any noteworthy correlation in terms of sign, significance and magnitude. Therefore, we list the correlations between the variables for each country and refrain to report cross-country correlations. Therefore, we conclude that both measures for information demand, the change in ASVI based on search terms as well as the predefined real estate category, reflect distinct indicators for public attention in each country and are to be included in the further analysis.

Empirical Findings

The aim of this study is to analyze the content and diffusion of public information in the context of stock markets. We investigate listed real estate stocks as either the information supply by the media and information demand by individual investors are curtailable. We derive measures for the content of information supply from Thomson Reuters News Archive and apply sentiment analysis to transform the texts into quantifiable measures. The demand for information is obtained from web search queries which approximate the revealed attention of individuals to the real estate sector. Hence, we assume that the content and the diffusion of public information by the same token affect stock market volatility.

As we observe time-varying volatility in our sample data on stock returns, we presume conditional heteroscedasticity in the error term. It has been observed more than half a century ago that stock returns exhibit prolonged periods of high returns followed by periods of low returns (Mandelbrot 1963). In this case of clustering volatility, the application of Autoregressive Conditional Heteroscedasticity ARCH (Engle 1982) and Generalized ARCH (GARCH) models (Bollerslev 1986) have been proven powerful tools to analyze financial time series. While prior studies have largely addressed the relevance of information on conditional volatility based on proxies such as trading volume, the impact of news content and media demand on stock returns have been addressed only recently (Da et al. 2014; Tetlock 2007). Nevertheless, only a few prior studies elaborate on clustering volatility rather than stock returns (Kalev et al. 2004).

In a first step, we average the variables for information supply and information demand in case of non-consecutive trading days following the procedure of García (2013). That is, we aggregate the news content and search queries available prior to market opening. While the bulk of our news data matches the trading days in our sample, we identify news content in 329 (UK) and 381 (US) days during which the market was closed. Consequently, for h days such that $h>0$ during which the market is closed and the respective two non-consecutive trading days t and $t+h+1$, we define the Negativity measure as

$$Neg_t^d = \frac{\sum_{i,s=t}^{s=t+h} n_{it}^d}{\sum_{i,s=t}^{s=t+h} w_{it}^d}$$

and proceed accordingly for the news sentiment measure ($Sent_t^d$) according to

$$Sent_t^d = \frac{\sum_{i,s=t}^{s=t+h} p_{it}^d - \sum_{i,s=t}^{s=t+h} n_{it}^d}{\sum_{i,s=t}^{s=t+h} w_{it}^d}.$$

Similarly, the time series reflecting the revealed attention of individuals to real estate content are converted into averages according to $\Delta ASVI_t^c = \sum_{j,s=t}^{s=t+h} \Delta ASVI_{j,t}^c$.

We employ a GARCH(1,1) model to investigate whether and how news and web searches as proxies for information impact the volatility of stock returns. The model considers a measure of conditional volatility taking time-varying second-order moments into account and is partly expressed in the form of

$$R_t^C = \mu SR_t^C + \rho + \varepsilon_t$$

as the mean equation, modeling the returns of the REIT indices R_t^C as a function of stock index returns SR_t^C, a constant mean ρ and an error term ε_t. The preconditions for the applicability of GARCH models are clustering volatility and ARCH effects in the residual. We approximate the mean equation, employ a Lagrange Multiplier test for autocorrelation and conclude that the mean models for the UK and the US have clustering volatility and ARCH effects in the residuals.

In order to investigate the effect of information flow on stock return volatility, we augment a GARCH(1,1) model with measures for information supply and information demand. That is, we base our analysis on the variables in the equation of conditional volatility in the form of

$$\sigma_t^2 = \omega + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2.$$

In this setting, the first order ARCH term is represented by the squared residual of the previous day (ε_{t-1}^2) and the first order GARCH term by the residual variance of the previous trading day ($\sigma_{t-1}^2 = var(\varepsilon_{t-1})$). The variance equation is characterized by a typical Autoregressive-Moving Average (ARMA) structure in which the coefficients α and β shed light on the dependence of current volatility on previous levels. Further, the

sum of the coefficients reflects to which degree volatility is persistent. We augment this setting by individually adding the media measures *Sent* and *Neg* and web search behavior *ASVI* as proxies for information supply *IS* and information demand *ID* in the form of

$$\sigma_t^2 = \omega + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2 + \lambda IS_t + \chi ID_t.$$

The empirical results for the UK are reported in Table 4 and reveal a significant impact of information demand and information supply on the volatility of REIT returns. The results of the GARCH(1,1) model including the information supply measures (IS_t^C) indicate a highly significant coefficient λ which is positive for Neg and negative for $Sent$ measures. Interestingly, news measures based on negative content (Neg) have a clear positive impact on the conditional volatility of REIT returns while the effect for net sentiment measures ($Sent$) exhibit contrary signs. The coefficients with p-values in paranthesis indicate a high level of significance.

Specification	Estimates of Parameters for UK REIT Returns							Diagnostics	
	ϱ	μ	ω	α	β	$\alpha+\beta$	λ	χ_t	LR Test
Without News	0.068	0.064	-0.579	0.252	0.902	1.154			
	(0.005)	(0.031)	(0.000)	(0.000)	(0.036)				
News Content	0.052	0.058	-1.020	0.304	0.557	0.861	0.805		2.22x10²
(Neg GI)	(0.033)	(0.073)	(0.000)	(0.000)	(0.000)		(0.000)		(0.032)
News Content	0.050	0.051	-1.890	0.318	0.632	0.950	0.570		1.95x10²
(Neg HE)	(0.048)	(0.113)	(0.002)	(0.000)	(0.000)		(0.000)		(0.048)
News Content	0.053	0.058	-0.320	0.228	0.467	0.695	0.902		3.02x10²
(Neg LM)	(0.041)	(0.079)	(0.039)	(0.000)	(0.000)		(0.000)		(0.040)
News Content	0.057	0.054	-1.425	0.317	0.593	0.910	-0.781		2.10x10²
(Sent GI)	(0.019)	(0.091)	(0.001)	(0.000)	(0.000)		(0.000)		(0.019)
News Content	0.047	0.049	-1.011	0.315	0.546	0.861	-0.921		2.21x10²
(Sent HE)	(0.064)	(0.126)	(0.000)	(0.000)	(0.000)		(0.000)		(0.064)
News Content	0.053	0.053	-0.523	0.249	0.501	0.750	-0.940		2.73x10²
(Sent LM)	(0.035)	(0.109)	(0.004)	(0.000)	(0.000)		(0.000)		(0.034)
Web Searches	0.027	0.111	-3.316	0.153	0.844	0.988		-0.135	1.28x10³
(Search Terms)	(0.407)	(0.000)	(0.000)	(0.000)	(0.000)			(0.851)	(0.407)
Web Searches	0.063	0.054	-4.370	0.303	0.680	0.983		1.349	7.68x10²
(Search Category)	(0.018)	(0.111)	(0.076)	(0.000)	(0.000)			(0.079)	(0.0184)

Table 4. Effects of News Content and Web Search Queries on Conditional Volatility of UK REIT Returns.

In broad terms, bad news roughly increase while good news reduce conditional volatility. In addition, the persistence of conditional volatility significantly decreases with news content variables included in the model. We note differences between our news content measures. We observe that the inclusion of LM-based news content measures substantially reduces the volatility measures to a significantly higher extent than the other media measures. The likelihood ratio (LR) statistic indicates the appropriateness of the model in relation to the base model without the inclusion of any information proxy. The LR is calculated as twice the difference between the log likelihood value of each model specification and the base model.

Based on these results, we suggest that including the variables $Neg(LM)$ and $Sent(LM)$ provides the best fit in characterizing the relationship between information supply variables and stock market volatility in our sample data. The estimations of the GARCH(1,1) model including proxies for information demand (ID_t^C), however, do not exhibit comparable levels of significance. In fact, abnormal search behavior is only significant on the 10%-level. We additionally note just marginal reductions of volatility persistence regarding web search behavior.

Table 5 reports the estimated results for the mean and variance equations for US REIT returns. Similar to the evidence exhibited for the UK, the impact of the news content on the conditional volatility of stock returns depends on the information variables. News content measures based on negative connotations (Neg) have a strong positive coefficient while the impact of net sentiment measures ($Sent$) on conditional volatility is negative.

Specification	Estimates of Parameters US REIT Returns							Diagnostics	
	ϱ	μ	ω	α	β	$\alpha+\beta$	λ	χ_t	LR Test
Without News	1.106	0.050	0.162	0.520	0.417	0.937			
	(0.000)	(0.007)	(0.000)	(0.000)	(0.000)				
News Content	1.113	0.048	-1.710	0.525	0.397	0.922	0.272		2.57x10²
(Neg GI)	(0.000)	(0.018)	(0.000)	(0.000)	(0.000)		(0.008)		(0.000)
News Content	1.110	0.044	-1.650	0.515	0.390	0.905	0.384		2.66x10²
(Neg HE)	(0.000)	(0.021)	(0.000)	(0.000)	(0.000)		(0.000)		(0.000)
News Content	1.113	0.047	-1.280	0.526	0.327	0.853	0.484		2.83x10²
(Neg LM)	(0.000)	(0.017)	(0.000)	(0.000)	(0.000)		(0.000)		(0.000)
News Content	1.113	0.044	-1.600	0.525	0.380	0.905	-0.351		2.63x10²
(Sent GI)	(0.000)	(0.021)	(0.000)	(0.000)	(0.000)		(0.000)		(0.000)
News Content	1.112	0.045	-1.398	0.504	0.352	0.856	-0.614		2.85x10²
(Sent HE)	(0.000)	(0.019)	(0.000)	(0.000)	(0.000)		(0.000)		(0.000)
News Content	1.112	0.0471	-1.190	0.513	0.315	0.828	-0.524		2.90x10²
(Sent LM)	(0.000)	(0.018)	(0.000)	(0.000)	(0.000)		(0.000)		(0.000)
Web Searches	1.140	0.056	-1.896	0.558	0.410	0.968		0.3318	7.03x10²
(Search Terms)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)			(0.040)	(0.000)
Web Searches	1.522	-0.0148	-1.654	0.468	0.466	0.934		0.633	5.97 x10²
(Search Category)	(0.000)	(0.497)	(0.000)	(0.000)	(0.000)			(0.000)	(0.000)

Table 5. Effects of News Content and Web Search Queries on Conditional Volatility of US REIT Returns.

However, even after the inclusion of our information variables the ARCH and GARCH effects still persist in terms of significance. We assume that this is mostly due to the fact that the variables arguably bear public information and do not capture all relevant and in particular private information. However, the significant relationship of information supply and demand variables as well as the consistency among the measures for both countries provide evidence for the notion that investor psychology affects stock markets. More precisely, we find that information supply and information demand partly explain market volatility in real estate-related stock markets.

Conclusion

This study introduces and analyzes blue ocean IS measures of information supply and information demand in stock markets. The research is motivated by the increasing availability of an extensive volume, variety and velocity of information. While prior work adopting text mining and Big Data analytics has addressed the notion that the arrival of new information affects asset markets, an application in the securitized real estate domain is novel. Our empirical results reveal a significant influence of media content and search behavior on the conditional volatility of real estate investment trust (REIT) stock returns. We adopt news sentiment as a proxy for information supply and Google Trends data as an indicator for revealed information demand by investors. The estimation of a GARCH(1,1) model with public information flow on conditional volatility of real estate stock indicates that in particular the content of information is being processed by investors. The main finding of this study is an analogous impact of information in the United Kingdom and the United States. While sentiment measures based on negative connotations increase conditional volatility analogous to the negativity-bias emphasized in the psychology discipline (Baumeister et al. 2001), the additional inclusion of positive content reduces persistent stock price volatility. While revealed attention approximated through Google Trends data is found to be significant, information processing on the basis of news content seems to exhibit closer relations to asset price volatility. However, our approximations for information supply and information demand seem not to inhibit all relevant information. We argue that this is due to the nature of our measures, which take public but no private information into account. In order to augment the current state of research, future work could employ additional proxies for information and additionally model combined effects of information arrival.

REFERENCES

Akerlof, G. A., and Shiller, R. J. 2009. *Animal Spirits: How Human Psychology Drives the Economy*: *How Human Psychology Drives the Economy, and Why It Matters for Global Capitalism*, Princeton: Princeton University Press.

Antweiler, W., and Frank, M. Z. 2004. "Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards," *The Journal of Finance* (59:3), pp. 1259–1294.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., and Vohs, K. D. 2001. "Bad Is Stronger Than Good," *Review of General Psychology* (5:4), pp. 323–370.

Bollerslev, T. 1986. "Generalized autoregressive conditional heteroskedasticity," *Journal of Econometrics* (31:3), pp. 307–327.

Brooks, C., and Tsolacos, S. 2000. "Does Orthogonalization Really Purge Equity-Based Property Valuations of Their General Stock Market Influences?" *Applied Economics Letters* (7:5), pp. 305–309.

Choi, H., and Varian, H. 2012. "Predicting the Present with Google Trends," *The Economic Record* (88:1), pp. 2–9.

Da, Z., Engelberg, J., and Gao, P. 2011. "In Search of Attention," *The Journal of Finance* (66:5), pp. 1461–1499.

Da, Z., Engelberg, J., and Gao, P. 2014. "The Sum of All FEARS Investor Sentiment and Asset Prices," *Review of Financial Studies* (27:12), pp. 1–33.

DeBondt, W., and Thaler, R. H. 1985. "Does the Stock Market Overreact?" *Journal of Finance* (40:3), pp. 793–805.

Demers, E. A., and Vega, C. 2010. "Soft Information in Earnings Announcements: News or Noise?" *SSRN Electronic Journal* .

Drake, M. S., Roulstone, D. T., and Thornock, J. R. 2012. "Investor Information Demand: Evidence from Google Searches Around Earnings Announcements," *Journal of Accounting Research* (50:4), pp. 1001–1040.

Engle, R. F. 1982. "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kindom Inflation," *Econometrica* (50:4), pp. 987–1007.

Esuli, A., and Sebastiani, F. 2006. "SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining," in *Proceedings of 5th International Conference on Language Ressources and Evaluation (LREC)*, pp. 417–422.

García, D. 2013. "Sentiment during Recessions," *The Journal of Finance* (68:3), pp. 1267–1300.

Grefenstette, G., and Tapanainen, P. "What is a Word, What is a Sentence? Problems of Tokenization," in *Proceedings of the 3rd Conference on Computational Lexicography and Text Research*, pp. 78–87.

Henry, E. 2008. "Are Investors Influenced By How Earnings Press Releases Are Written?" *Journal of Business Communication* (45:4), pp. 363–407.
Henry, E., and Leone, A. J. 2009. "Measuring Qualitative Information in Capital Markets Research," *SSRN Electronic Journal* .
Hirshleifer, D. 2001. "Investor Psychology and Asset Pricing," *Journal of Finance* (56:4), pp. 1533–1597.
Hu, M., and Liu, B. 2004. "Mining Opinion Features in Customer Reviews," *Proceedings of the 19th national conference on Artifical intelligence (AAAI)* , pp. 755–760.
Jegadeesh, N., and Di Wu 2013. "Word power: A new approach for content analysis," *Journal of Financial Economics* (110:3), pp. 712–729.
Kahneman, D., and Tversky, A. 1979. "Prospect Theory: An Analysis of Decision under Risk," *Econometrica* (47:2), pp. 263–292.
Kalev, P. S., Liu, W.-M., Pham, P. K., and Jarnecic, E. 2004. "Public information arrival and volatility of intraday stock returns," *Journal of Banking & Finance* (28:6), pp. 1441–1467.
Keynes, J. M. 1936. *The General Theory of Employment, Interest and Money*, London.
Loughran, T., and McDonald, B. 2011. "When Is a Liability Not a Liability Textual Analysis, Dictionaries, and 10Ks," *Journal of Finance* (66:1), pp. 35–65.
Loughran, T., and McDonald, B. 2013. "IPO first-day returns, offer price revisions, volatility, and form S-1 language," *Journal of Financial Economics* (109:2), pp. 307–326.
Lovins, J. B. 1968. "Development of a Stemming Algorithm," *Mechanical Translation and Computational Linguistics* (11:1), pp. 22–31.
Mandelbrot, B. 1963. "The Variation of Certain Speculative Prices," *The Journal of Business* (36), p. 394.
Manning, C. D., and Schütze, H. 1999. *Foundations of Statistical Natural Language Proccesing*, Cambridge (Massachusetts): MIT Press.
McCue, T. E., and Kling, J. L. 1994. "Real Estate Returns and the Macroeconomy. Some Empirical Evidence from Real Estate Investment Trust Data," *Journal of Real Estate Research* (9), pp. 277–287.
Mohammad, S., and Turney, P. 2010. "Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon," *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text* , pp. 26–34.
Morawski, J., Rehkugler, H., and Füss Roland 2008. "The Nature of Listed Real Estate Companies: Property or Equity Market?" *Financial Markets and Portfolio Management* (22:2), pp. 101–126.
Pang, B., and Lee, L. 2008. "Opinion Mining and Sentiment Analysis," *FNT in Information Retrieval (Foundations and Trends in Information Retrieval)* (2:1-2), pp. 1–135.
Porter, M. F. 1980. "An Algorithm for Suffix Stripping," *Program* (14:3), pp. 130–137.
Preis, T., Moat, H. S., and Stanley, H. E. 2013. "Quantifying Trading Behavior in Financial Markets Using Google Trends," *Scientific Reports* (3:1684).
Stiglitz, J. E. 2009. "The Global Crisis, Social Protection and Jobs," *International Labour Review* (1), pp. 1–12.
Stone, P. J. 1968. *The General Inquirer: A Computer Approach to Content Analysis*, Cambridge (Massachusetts): MIT Press.
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.
White, H. 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica* (48:4), pp. 817–838.
Wilson, T., Wiebe, J., and Hoffmann, P. 2005. "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis," *Proceedings of Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)* , pp. 347–354.

Acknowledgements

The author gratefully acknowledges the support of Martin Dixon and Franz Fuerst at the Department of Land Economy while being a visiting scholar at the University of Cambridge. The author is indebted to the anonymous referees for valuable suggestions. The responsibility for all remaining errors is, of course, mine.