Evaluating the impact of social-media on sales forecasting: a quantitative study of worlds biggest brands using Twitter, Facebook and Google Trends.

Olga Kolchyna

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
of
University College London.

Department of Computer Science
University College London

December 19, 2017

I, Olga Kolchyna, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

In the world of digital communication, data from online sources such as social networks might provide additional information about changing consumer interest and significantly improve the accuracy of forecasting models. In this thesis I investigate whether information from Twitter, Facebook and Google Trends have the ability to improve daily sales forecasts for companies with respect to the forecasts from transactional sales data only. My original contribution to this domain, exposed in the present thesis, consists in the following main steps:

1. Data collection. I collected Twitter, Facebook and Google Trends data for the period May 2013 May 2015 for 75 brands. Historical transactional sales data was supplied by Certona Corporation.

2. Sentiment analysis. I introduced a new sentiment classification approach based on combining the two standard techniques (lexicon-based and machine learning based). The proposed method outperforms the state-of-the-art approach by 7% in F-score.

3. Identification and classification of events. I proposed a framework for events detection and a robust method for clustering Twitter events into different types based on the shape of the Twitter volume and sentiment peaks. This approach allows to capture the varying dynamics of information propagation through the social network. I provide empirical evidence that it is possible to identify types of Twitter events that have significant power to predict spikes in sales.

4. Forecasting next day sales. I explored linear, non-linear and cointegrating relationships between sales and social-media variables for 18 brands and showed that social-media variables can improve daily sales forecasts for the

majority of brands by capturing factors, such as consumer sentiment and brand perception. Moreover, I identified that social-media data without sales information, can be used to predict sales direction with the accuracy of 63%.

The experts from the industry consider the results obtained in this thesis to be valuable and useful for decision making and for making strategic planning for the future.

Impact Statement

This thesis is an empirical study investigating the role of social media for sales forecasting for retail companies. By combining two areas of research, event study and forecasting, this research expands existing literature on social media analysis opening up exciting opportunities for academic and industrial enquiry.

The findings of the thesis demonstrated that incorporation of user-generated data from Twitter, Facebook and Google Trends into forecasting models allows to increase the accuracy of forecasts for volume and direction of sales, as well as provides a possibility of predicting spikes in sales. These research outcomes already started to impact organisations by motivating managers to re-evaluate their business strategies to capitalise on the findings and gain competitive advantage.

My research has already been recognised by Certona, a leader in the market in providing recommendation platforms to world's biggest brands. Certona is in the process of integrating my findings and developing a new value proposition to its clients. The findings of my research could therefore help the managers of many retail brands to make better decisions regarding their marketing strategies, building customer relationships, planning inventory and performing sales forecasting making the system more efficient.

I have written four papers related to my work and presented my findings locally at a number of UCL events, nationally at the CFE-CMStatistics-2015 conference, at Sentiment Analysis Applied to Finance conferences in 2014 and 2015, as well as internationally at the International Conference on Computational Social Science 2015 in Helsinki, and at the Future Technologies Conference 2016 in San Francisco. I also conducted two workshops at Certona, using my research findings to extend the knowledge of Certona's data analytics team and demonstrating the value that social media can provide to the management of the company.

Dedicated to my parents, Natallia and Sergej.

Acknowledgements

I would like to thank my principal supervisor Prof. Tomaso Aste for his wisdom, time and patience, and helping me through every step of this PhD journey. While providing constructive feedback and giving advice, Tomaso also gave me the space to explore different ideas and be creative. The lessons that I learned from working under his guidance will stay with me for lifetime. I feel exceedingly appreciated to have had his support and unwavering encouragement, and I owe him a great many heartfelt thanks.

I would also like to thank my supervisor Prof. Philip Treleaven for providing a vision and helping me to see a big picture. His advice helped me to stay on track and to maintain positive attitude.

My appreciation goes to Prof. Henrik Jeldtoft Jensen and Prof. Tobias Preis for being the examiners of this thesis.

Thanks to the Engineering and Physical Science Research Council (EPSRC) for having funded my research.

I would like to acknowledge Certona Corporation for providing sales data for brands. I am particularly grateful to Dr. Geoffrey Hueter and Dr. Stephen Kerr for their consistent support and time taken to contribute their domain knowledge and expertise. Their suggestions and input were very helpful and allowed to produce the outcomes directly applicable for solving real-life business problems.

Many thanks to Giuseppe Pappalarado for collecting and providing Facebook data.

I would also like to thank Prof. Didier Sornette for his valuable and constructive suggestions during the development of this research work.

On a personal level, there are several people that deserve acknowledgement. I am grateful for the support of my parents and my sister, for their love and

unlimited support in helping me to pursue and achieve my goals. I would like to thank Denis Wallez for his friendship, for reviewing this work, providing valuable suggestions and encouraging me. My deep appreciation goes to my friend, Tatyana Kovalenko, who was also walking the PhD path at the same time as me, and shared with me all the ups and downs. And finally, my deepest gratitude is reserved for my partner, Rodrigo Mazorra Blanco, who supported me on a daily basis and especially, during the difficult moments. Thank you, Rodrigo, for providing me with encouragement and sharing your indomitable spirit.

Contents

Acronyms

ACF Autocorrelation Function

ADF Augmented Dickey-Fuller test

AIC Akaikes Information Criterion

ANN Artificial Neural Network

API Application Programming Interface

AR Auto-Regression

ARIMA Autoregressive Integrated Moving Average

AUTO Automatically Constructed Lexicon

BOW Bag-of-Words

DF Dickey-Fuller test

DM-test Diebold-Mariano test

DTW Dynamic Time Warping

EA Earnings Announcements

ECM Error Correction Model

ED Euclidean Distance

ELM Extreme Learning Machine

EMO Emoticons Lexicon

ESD Extreme Studentized Deviation

FWER Family-Wise Error Rate

HDFS Hadoop Distributed File System

IDF Inverse Term Frequency

IQR Median and InterQuartile Range

KDE Kernel Density Estimation

KPSS Kwiatkowski-Phillips-Schmidt-Shin test

LR Logistic Regression

MAD Median Absolute Deviation

MAE Mean Absolute Error

MAPE Mean Absolute Percentage Error

MdAPE Median Absolute Percentage Error

MLR Multiple Linear Regression

MPQA Multi-Perspective-Question-Answering

OL Opinion Lexicon

OLS Ordinary Least Squares

POS Part-of-Speech

RelMAE Relative Mean Absolute Error

REST Representational State Transfer

SemEval Semantic Evaluation

SMIA Social Media Impact Analysis

SVM Support Vector Machine

TF Term Frequency

VAR Vector Auto-Regression

VECM Vector Error Correction Model

List of Figures

List of Tables

Chapter 1

Introduction

This chapter presents a brief overview of the development of social media and their implications on today's business operations. The motivations for this thesis are discussed and the main objectives are identified. The chapter also describes the main contributions and the structure of the thesis.

1.1 The Value of Social Media Data for Brands

During the last decade the media has experienced a huge transformation and extended beyond its original applications. With rapid spread of the Internet, people are turning away from traditional media, such as newspapers, radio and TV, and are increasingly using digital social media in search of information, such as microblogging sites, social networks, forums and virtual worlds (Schivinski and Dabrowski, 2016). These social media technologies have revolutionized the role of consumers, changing them from being passive observers and receivers of information into active participants, who interact, share knowledge and collaborate with each other. For example, Facebook and Google+ became highly popular between friends and family, allowing to share information through pictures and status updates, creating group chats for discussions and organising events. Other social networks like Tumblr or Twitter were designed for rapid communication and have become valuable means for spreading breaking news or providing timely and fine-grained information about important events as they unfold, while reflecting personal perspectives, emotional reactions, and controversial opinions.

Recent statistics indicates that there are over 2.80 billion global social media users (WeAreSocial, 2017), with Facebook being the most popular social network, reaching 2 billion active users in August 2017 (Statista, 2017). The amount of data that active users generate is enormous and can be described by a term ”big data”. It is becoming increasingly evident that by surveying and analysing this abundant and continual flow of user-generated content (collective intelligence) it is possible to understand human behaviour and predict different social phenomena, from unemployment rates (Antenucci et al., 2014) and influenza infection rates (Culotta, 2013, St Louis and Zorlu, 2012, Paul and Dredze, 2011, Aramaki et al., 2011) to election results (Bermingham and Smeaton, 2011, Tumasjan et al., 2010, Kim and Hovy, 2007).

One area that has the greatest potential to be dramatically changed by the technological advancements of social media is the area of business. For example, the adoption of social networks has replaced many traditional one-to-many business models with one-to-one approaches (Ngai et al., 2015). Today companies practice direct communication with its customers to address any issues related to their products or enumerate its loyal customers with gifts and promotions. According to Facebook, more than 50 million businesses are using Facebook pages in order to interact with their fans (Cohen, 2015). Gamboa and Gonalves (2014) demonstrated that Facebook is instrumental in achieving customer loyalty by increasing customer satisfaction and perceived value. A survey by Ambassador (2013) revealed that consumers expect to receive assistance within 5 minutes of reaching out to a company, and 71% of consumers who had a good social media interaction with a company are likely to recommend it to their friends and family. This statistics demonstrates that one-to-one approach is the future of building and maintaining brand reputation and obtaining new customers.

The way that companies perform marketing and branding is also changing. Historically, most of the money spent for advertising is wasted, but because user generated content provides an immediate feedback from people regarding new product lines, brand policies and actions, today companies can launch a campaign and measure its effectiveness through Twitter and other social media data almost in real time. The managers of a company can now incorporate the findings of the analytics reports to adjust marketing strategies with regards to the brand perception and consumer needs.

While the processes of branding, marketing, consumer engagement have been greatly impacted by social media innovations, the value of user generated data for businesses can be the most tangible for the purpose of understanding and predicting sales performance. Today, consumers are used to leaving feedback about their customer experiences and express views about products on social media websites, at the same time, people who are interested in purchasing a product are going online to read reviews before making a decision to buy. The study by Hudson and Thal (2013) analysed purchase behaviour of nearly 20,000 consumers across five industries and three continents, and observed that there were four stages of a consumer journey: (a) stage of consideration; (b) stage of evaluation; (c) stage of buying; and (d) stage of enjoying, advocating and bonding. The researchers highlighted that during the ”evaluation” and ”advocating” stages the impact of social media was extremely important. According to another study of 500 individuals, conducted by Shoutly [1] (Shoutly, 2014), 92% of respondents choose to read product recommendations before buying. 75% of customers say they use social media as part of the buying process (Pick, 2015b), while as many as 53% of Twitter users recommend products or brands themselves, and 48% of people decide to buy the product following the recommendations (Flannagan, 2014). As numbers indicate, the decision which product or service to choose is being greatly affected by other people’s feedback. By expressing their viwes online people set up trends and sentiments in the market. As Wright (2009) said, ”for many businesses online opinion has turned into a kind of virtual currency that can make or break a product in the marketplace”.

As consumers continue to rely on personal recommendations through social media, it has become essential for companies to start collecting and analysing online data of reviews, sentiments, comments, post shares and likes in order to understand customer purchase decisions and sales performance, so that the management can manage its supply chain more productively (Chong et al., 2017a). Organisations that succeed in incorporating the vast amount of web generated information into their models for demand forecasting and inventory management could gain a strong competitive advantage over their rivals (Chong and Li, 2014). The purpose of this study is to perform a large empirical investigation of the

[1] Shoutly is an online social commerce platform: https://shoutly.com

value of social media for sales forecasting for retail brands. The evidence of forecasts being improved by social media variables would encourage managers to change their business strategies and incorporate frameworks for data collection and analysis in the business pipelines, aiming at incremental benefits as brand equity, brand awareness and revenue potential.

1.2 Motivation and Objectives of this Thesis

The traditional way of predicting purchase behaviour and building forecasting models is mainly based on using historical sales data, customer related data from CRM, marketing and survey data (Berbegal-Mirabent et al., 2016). The decades of such data makes business managers confident in making predictions, pricing and taking other decisions regarding the customer base (Bradlow et al., 2017). However, in the world of social networks, market trends and sentiments change rapidly, making it questionable whether old historical data remains relevant for understanding current consumer demand and sales dynamics (Gur Ali and Pinar, 2016). In fact, old historical data that does not contain recent exogenous indicators of consumer sentiment can be misleading and result in poor decisions.

Differently from traditional approaches, the motivation of this study is to utilise rich data from online sources in order to capture early signals of changing consumer demand and, thus, improve the accuracy of forecasting models. When choosing the types of online data to be used for examining the effect on sales, most research efforts in recent years have been focused on the anonymous online reviews. While many studies discovered a positive relationship between reviews and sales performance (Chong et al., 2017b, Bao and Chang, 2016, Schneider and Gupta, 2016, Chong et al., 2016, Bao and Chang, 2014, Liu et al., 2010, Chevalier and Mayzlin, 2006), some researchers found no or limited effect of ratings of reviews on sales (Davis and Khazanchi, 2008, Wenjing et al., 2008, Amblee and Bui, 2007, Hu et al., 2006, Chen et al., 2004). These contradictory findings might be caused by the systematic biases related to the reviewing process: it has been shown that later reviews are often being influenced by prior reviews (Ma et al., 2014). Inconsistent results of reviews impact on sales might also be related to the anonymous nature of online reviews. Mayzlin (2006) showed that many companies exploit the popularity of Internet and fabricate positive reviews related to their brand, pretending to be consumers, and this way they make online reviews less reliable.

While the number of people who trust opinions of anonymous users is high (70%), even more consumers trust recommendations from people they know (90%) (Flannagan, 2014). Social media platforms that require authorisation of users offer a fundamental advantage over anonymous systems as they present opinions of friends and family rather than unknown people (Lee et al., 2016). A public survey by Shoutly (2014) revealed that the most trusted social media platforms for product recommendations are Facebook (65%), followed by YouTube (36%) and then Twitter (27%). In this thesis, I used Twitter and Facebook data as sources of consumers' opinions, being motivated to avoid biases related to anonymous reviews. I also analysed Google Trends data that represents the number of Google searches. Since 89% of customers begin their buying process with a search engine (Pick, 2015a), Google Trends data may present an indication of consumers' purchasing intent and can be extremely relevant for future sales prediction.

This thesis sets out to achieve three main objectives:

1. **Measure the ability of Twitter events to predict spikes in sales.**

   The possibility to foresee spikes in sales is a key success factor for any business as information about changing demand would allow to perform accurate replenishments and inventory allocation. The first objective of this thesis is based on the hypothesis that spikes in sales can be predicted by a sudden change of brand-related sentiment on social media, particularly Twitter. This hypothesis is inspired by the previous research that showed how events in Twitter can be used to predict abnormal market returns (Ranco et al., 2015, Sprenger et al., 2014), riot events (Alsaedi et al., 2017), protests (Kallus, 2014), etc. However, little is known about the ability of social media events to predict spikes in sales of a company. To my best knowledge, only one study performed analysis of Twitter events in relation to sales (Dijkman et al., 2015). Dijkman et al. (2015) showed that spikes in volume of positive tweets were significantly correlated with sales of the next few weeks. However, the results of this study are hard to generalise since the analysis was performed only for 12,780 tweets related to one company. Chapter 5 of this thesis performs analysis of sales and Twitter events for 75 brands and aspires to address this research gap.

2. **Evaluate the effects of Twitter, Facebook and Google Trends on short term sales forecasting.**

   Many researchers have demonstrated the value of social media for improving sales forecasts (Liu, 2006, Lassen et al., 2014, Kulkarni et al., 2012, Asur and Huberman, 2010, Gruhl et al., 2005). However, existing studies predominantly explored sales for products that receive a lot of social attention, such as movies, books, iphones. ”General” or ”non-trending” types of products were studied only by a few researchers (Lee et al., 2016, Boldt et al., 2016, Du et al., 2015), and their analysis is limited to considering only quarter or annual sales.

   The motivation behind the second objective of this thesis is to evaluate whether social media data can improve daily, rather than quarter or annual, sales forecasts for ”general” consumer products. There is a large gap in the literature related to the scarcity of studies on daily sales forecasting, caused by a fact that sales data is a very sensitive and protected piece of information for any company. The only data that is generally available to researchers is revenues extracted from quarter or annual financial reports. While previous studies succeeded to demonstrate the impact of social media data on annual sales, the question regarding the short term value of user-generated content remains open. The hypothesis of the current research work is that social media can provide timely information about changing consumer demand. This information can be extremely valuable for the management, allowing them to adjust value proposition in respond to changing demand almost in real time, or evaluate the outcomes of new product releases and marketing campaigns within hours or days from launching. To test the hypothesis and address the gap in the literature I collaborated with a company Certona, which was interested in exploring the role of social media for brands and provided transactional sales data of its clients.

3. **Develop a framework for evaluating the impact of social media on sales**

and generalise the findings across multiple brands and multiple social media sources.

Previous research on the role of social media has only covered few types of products and few companies. The third objective of this thesis is to generalise previous findings by studying 75 brands from the retail sector, including apparel, footwear, accessories, body care, furnishings, equipment and games. Furthermore, existing research works on the impact of social media on sales only considered one social media indicator at a time. This thesis aims to extend previous works by incorporating multiple types of online data, such as Twitter, Facebook, and Google Trends. Moreover, the motivation behind this study is to develop a framework that allows to collect social media data, automatically detect anomalous events, incorporate different social media indicators into forecasting techniques, evaluate and select the best performing models.

1.3 Case Study: Certona

This research was done in collaboration with the company Certona. Certona is the leader in providing personalized experiences for the world's most popular brands, such as Nike, Adidas, Denim, Nikon among many others. Certona's personalization platform collects real-time data about the consumers browsing actions from more than four hundred top e-commerce sites, and uses this data for real-time behavioural profiling, serving up individualized content, promotional and product recommendations. The main goal of the company is to personalize the entire visitor experience for each individual, including web experience, e-mail, mobile, in-store, call center and social networks, which results in increased revenues and customer loyalty.

Certona recognizes the growing role of social media for brands. One of the priorities that Certona has today is to analyse what people say about different brands, how their sentiment towards a brand is changing and why. Collaboration with Certona on this thesis allowed to align the research objectives with the real business needs, perform a real-life quantitative study and receive feedback on the value of the findings from business leaders.

1.4 Thesis' Contributions

This study has made several practical, theoretical and methodological contributions.

Practically:

- This study has practical implications to managers and practitioners: it demonstrates that user-generated data, such as Twitter messages, Facebook posts and likes, and Google searches can be used to forecast the volume and direction of sales for retail brands, and, in some cases, predict spikes in sales. Following these findings business managers might be motivated to develop new strategies that incorporate social media data into their every-day decision making process related to inventory management, demand forecasting, consumer analysis, and marketing planning.
- The study presents a large-scale analysis of 75 retail brands, allowing to generalise the findings across multiple sectors: apparel, footwear, accessories, body care, furnishings, equipment and games.
- Previous research works focused on the analysis of one social media indicator at a time. This thesis compares performance of Twitter, Facebook and Google Trends, and shows that Google Trends is the strongest indicator of changing consumer demand, followed by Facebook and then Twitter.
- The majority of the previous studies analysed quarter or annual sales, leaving a gap in the literature related to daily sales forecasting. This thesis fills in the gap by evaluating the value of social media data for predicting next-day sales. The findings demonstrate that social media data can provide timely information about changing consumer demand and improve next-day sales forecasts. This should give managers an incentive to seriously consider social media sentiment when measuring marketing campaigns and analysing the performance of new product lines.
- This thesis provides a detailed description of the data retrieval, pre-processing, storing and forecasting steps that comprise a Social Media Impact Analysis ('SMIA') framework. This framework can be directly adopted by industry leaders within the retail domain or any other sector impacted by social media.

Theoretically:

- While the previous studies have demonstrated that social media data can be used to predict different social, political and economic phenomena, such as riot events, protests, stock markets, elections, very limited research has been done on analysis of events in social networks for predicting spikes in sales. This thesis provides a contribution to event study literature by suggesting the approach for detecting events in Twitter and measuring their ability to predict spikes in sales.

- The differentiation between social media events in existing studies has primarily been done based on spacial, textual and temporal information. In this research work I suggest to cluster Twitter events based on the shape of the event's growth and relaxation signatures. I show that this approach captures different dynamics of information dissemination in Twitter and allows to identify classes of events that have significant probability to be followed by spikes in sales.

Methodologically:

- This thesis presents a contribution to the event study field by proposing a new events clustering method based on slopes of growth and relaxation signatures.

- The traditional way of defining an event is detecting a peak and selecting an event window. The process of selecting the optimal event window presents a challenge. This study contributes to the methodology of event identification by proposing an approach for automatic detection of the optimal event window.

- The findings of this thesis also extend the field of sentiment analysis by suggesting a hybrid method that combines a lexicon-based approach with a machine learning approach, and leverages the "social media language": emoticons, slang and abbreviations. A proposed method outperformed a state-of-the-art method by 4%.

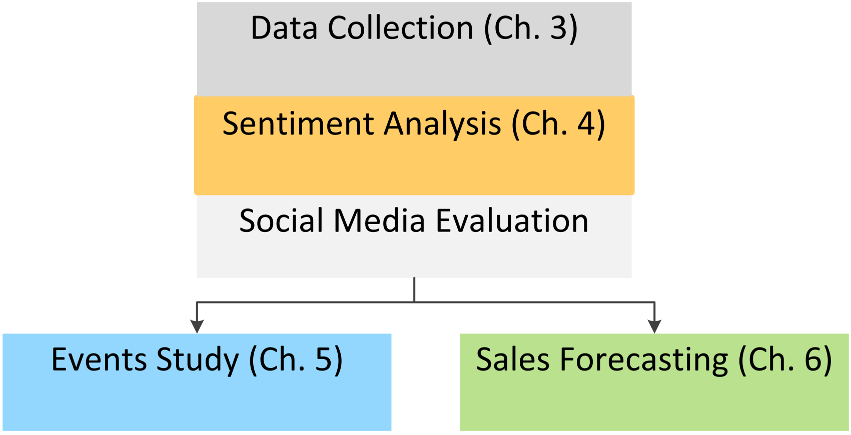


Figure 1.1: Schematic representation of the steps performed in this thesis.

- Finally, the analytics framework described in this thesis suggests a methodological approach for stationarity and cointegration testing, as well as training and selecting models that are most suitable for the types of variables under analysis.

1.4.1 Publications

The work presented in this thesis resulted in the following publications:

Kolchyna, O., Souza, T. T. P., Treleaven, P. C., and Aste, T. ”Twitter Sentiment Analysis: Lexicon Method, Machine Learning Method and Their Combination”, In Handbook of Sentiment Analysis in Finance, Mitra, G. and Yu, X. (Eds.), chapter 5, 2016

Souza, T. T. P., Kolchyna, O., Treleaven, P. C., and Aste, T. ”Twitter Sentiment Analysis Applied to Finance: A Case Study in the Retail Industry”, In Handbook of Sentiment Analysis in Finance, Mitra, G. and Yu, X. (Eds.), chapter 23, 2016

Kolchyna, O., Souza, T. T. P., Treleaven, P. C., and Aste, T. ”A framework for twitter events detection, differentiation and its application for retail brands.” In 2016 Future Technologies Conference (FTC), IEEE Xplore, pages 323 - 331, 2016

Kolchyna, O., Treleaven, P. C., and Aste, T. (2017). Estimating daily sales for brands from Twitter messages, Facebook posts and Google Trends. Submitted to Decision Support Systems Journal.

1.5 Structure of the Thesis

The remainder of this thesis is structured as follows:

- Chapter 2 presents background information, related works discussion and review of the most popular techniques for 4 fields of study used in this thesis: 1) sentiment analysis; 2) time series analysis; 3) events study; 4) time series forecasting. A reader might choose to skip this chapter and refer to specific methods or literature review when required, while reading the following chapters of the thesis.

- Chapter 3 is dedicated to the description of the social media data collection process. In this chapter the basics of Twitter API, Facebook API, Google trends API are discussed along with data filtering and pre-processing methods.

- Chapter 4 focuses on the sentiment analysis of the collected Twitter data. Data pre-processing is covered and two most popular methods for sentiment analysis: lexicon-based and machine learning approaches. Under the Lexicon-based approach the process of the lexicon creation and enhancement is described in detail. Under the Machine Learning approach a few algorithms are tested: Naive Bayes, Decision Trees and Support Vector Machines.

- Chapter 5 aims to achieve the first objective of the thesis by evaluating the relationship between Twitter events and sales events. The chapter discusses the methods for events detection, proposes a new method for events clustering, and uses two statistical tests to quantify the probability of having sales events after Twitter events.

- Chapter 6 aims to achieve the second objective of the thesis by performing sales forecasting using Twitter, Facebook and Google Trends data. Stationarity and cointegration tests of time series are performed in this chapter and different types of forecasting models are tested (multiple linear regression, artificial neural networks, vector error correction model).

- Chapter 7 contains the conclusions of the work, discusses its limitations, and suggests the future work that can be done in this area.

- The thesis finishes with the Appendices and a list of references.

Fig. 1.1 presents a high level of the ’SMIA’ framework used in this thesis. Every step in the framework corresponds to one of the chapters (3-6) and is denoted by a color for the ease of identification. Within the chapters more detailed frameworks are presented, marked by a corresponding color. The combination of the frameworks from each chapter comprises the ’SMIA’ framework that can be adopted by academics or practitioners, who would like to measure the impact of social media and incorporate it into their forecasting models.

Chapter 2

Background and Literature Review

The purpose of this chapter is to introduce the key concepts and methods of sentiment analysis, event detection and time series forecasting, that will be used in subsequent chapters of this thesis. The chapter also presents the state-of-the-art literature review for each of the topics.

2.1 Sentiment Analysis

2.1.1 Sentiment Analysis: Related Work

The field of text categorization was initiated long time ago (Salton and McGill, 1983), however categorization based on sentiment analysis was introduced more recently in Das and Chen (2001), Morinaga et al. (2002), Pang et al. (2002), Tong (2001), Turney (2002), Wiebe (2000).

Sentiment analysis is a line of research that allows to determine people's attitude and opinions in relation to different topics, products, services, people, events, and their attributes. As Hu and Liu (2004) highlight, in the research literature it is possible to see many different names, e.g. *"sentiment analysis, opinion mining, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, review mining"*, however, all of them have similar purposes and belong to the subject of sentiment analysis or opinion mining. The term sentiment analysis is more known in the industry than the term opinion mining.

The standard approach for text representation (Salton and McGill, 1983) has been the bag-of-words method (BOW). According to the BOW model, the document is represented as a vector of words in Euclidean space where each word is independent from others. This bag of individual words is commonly called a collection of unigrams. The BOW is easy to understand and allows to achieve high performance (for example, the best results of multi-lable categorization for the Reuters-21578 dataset were produced using BOW approach (Dumais et al., 1998, Weiss et al., 1999)).

The main two methods of sentiment analysis, lexicon-based method (Taboada et al., 2011, Ding et al., 2008) and machine learning based method (Pak and Paroubek, 2010), both rely on the BOW. In the machine learning supervised method the classifiers are using the unigrams or their combinations (n-grams) as features. In the lexicon-based method the unigrams which are found in the lexicon are assigned a polarity score, the overall polarity score of the text is then computed as sum of the polarities of the unigrams.

When deciding which lexicon elements of a message should be considered for sentiment analysis, different parts-of-speech were analysed (Pak and Paroubek, 2010, Kouloumpis et al., 2011). Benamara et al. (2007) proposed the Adverb-Adjective Combinations approach that demonstrates the use of adverbs and adjectives to detect sentiment polarity. In recent years the role of emoticons has been investigated (Pozzi et al., 2013, Hogenboom et al., 2013, Liu et al., 2012, Zhao et al., 2012, Go et al., 2009). Fersini et al. (2015) further explored the use of (i) adjectives, (ii) emoticons, emphatic and onomatopoeic expressions and (iii) expressive lengthening as expressive signals in sentiment analysis of microblogs. They showed that the above signals can enrich the feature space and improve the quality of sentiment classification. Jurek et al. (2015) suggested to classify messages not only as positive, negative or neutral, but to include the intensity of the sentiment by utilising intensifier words, such as “very”, “quite”, “most”, etc. Muhammad et al. (2016) improved the accuracy of sentiment classification by accounting for contextual polarities of words and by enhancing a general purpose lexicon with domain specific knowledge.

Advanced algorithms for sentiment analysis have been developed (Jacobs, 1992, Vapnik, 1998, Basili et al., 2000, Schapire and Singer, 2000) that take into consideration not only the message itself, but also the context in which the message is published, who is the author of the message, who are the friends of the author, what is the underlying structure of the network. For instance, Hu et al. (2013) investigated how social relations can help sentiment analysis by introducing a sociological approach to handling noisy and short texts, Zhu et al. (2014) showed that the quality of sentiment clustering for Twitter can be improved by joint clustering of tweets, users, and features, Pozzi et al. (2013) looked at friendship connections and estimated user polarities about a given topic by integrating post contents with approval relations, You and Luo (2013) improved sentiment classification accuracy by adding a visual content in addition to the textual information, Aisopos et al. (2012) significantly increased the accuracy of sentiment classification by using content-based features along with context-based features, Saif et al. (2012) achieved improvements by growing the feature space with semantics features.

While many research works focused on finding the best features, some efforts have been made to explore new methods for sentiment classification. Wang et al. (2014) evaluated the performance of ensemble methods (Bagging, Boosting, Random Subspace) and empirically proved that ensemble models can produce better results than the base learners. Fersini et al. (2014) proposed to use Bayesian Model Averaging ensemble method which outperformed both traditional classification and ensemble methods. Carvalho et al. (2014) employed genetic algorithms to find subsets of words from a set of paradigm words that led to improvement of classification accuracy. Poria et al. (2015), Severyn and Moschitti (2015), Stojanovski et al. (2015) achieved improved classification accuracies by implementing deep convolutional networks. Some researchers combined machine learning techniques with lexicon based methods. For example, Mudinas et al. (2012) extracted sentiment words and used them as features in machine learning algorithm. Zhang et al. (2011) used sentiment words extracted with the help of lexicon approach to discover additional sentiment words by using Chi-square test. Sentiment polarity of newly discovered words was established through a classifier, which was trained using initial sentiment words. Appel et al. (2016) combined semantic rules, fuzzy sets, and enriched sentiment lexicon to create a hybrid approach that outperformed Naive Bayes and Maximum Entropy methods on a movies reviews dataset.

2.1.2 Data Pre-processing for Sentiment Analysis

Prior to applying any of the sentiment extraction methods, a common practice is to perform data pre-processing. Data pre-processing allows to produce higher quality text classification and reduce the computational complexity. Typical pre-processing procedure includes the following steps:

Part-of-Speech Tagging (POS). The process of part-of-speech tagging allows to automatically tag each word of text in terms of which part of speech it belongs to: noun, pronoun, adverb, adjective, verb, interjection, intensifier, etc. The goal is to extract patterns in text based on analysis of frequency distributions of these parts-of-speech. The importance of part-of-speech tagging for correct sentiment analysis was demonstrated by Manning and Schütze (1999). Statistical properties of texts, such as adherence to Zipfs law can also be used (Piantadosi, 2014). Pak and Paroubek (2010) analysed the distribution of POS tagging specifically for Twitter messages and identified multiple patterns. For instance, they found that subjective texts (carrying the sentiment) often contain more pronouns, rather than common and proper nouns; subjective messages often use past simple tense and contain many verbs in a base form and many modal verbs.

There is no common opinion about whether POS tagging improves the results of sentiment classification. Barbosa and Feng (2010) reported positive results using POS tagging, while Kouloumpis et al. (2011) reported a decrease in performance.

Stemming and lemmatisation. Stemming is a procedure of replacing words with their stems, or roots. The dimensionality of the BOW is reduced when root-related words, such as "read", "reader" and "reading" are mapped into one word "read". However, one should be careful when applying stemming, since it might increase bias. For example, the biased effect of stemming appears when merging distinct words "experiment" and "experience" into one word "exper", or when words which ought to be merged together (such as "adhere" and "adhesion") remain distinct after stemming. These are examples of over-stemming and under-stemming errors, respectively. Overstemming lowers precision and under-stemming lowers recall. The overall impact of stemming depends on the dataset and stemming algorithm. The most popular stemming algorithm is Porter stemmer (Porter, 1980).

Stop-words removal. Stop words are words which carry a connecting function in the sentence, such as prepositions, articles, etc. (Salton and McGill, 1983). There is no definite list of stop words, but some of the most common words include "the", "is", "at", "which" and "on". These words can be removed from the text before classification, since they have a high frequency of occurrence in the text, but do not affect the final sentiment of the sentence.

Negations Handling. Negation refers to the process of conversion of the sentiment of the text from positive to negative or from negative to positive by using special words: *"no","not","don't,"* etc. These words are called negations. Handling negation in the sentiment analysis task is a very important step as the whole sentiment of the text may be changed by the use of negation (for information on scope of negation see Councill et al. (2010)). The simplest approach to handle negation is to revert the polarity of all words that are found between the negation and the first punctuation mark following it. For instance, in the text "I don't want to go to the cinema," the polarity of the whole phrase "want to got to the cinema" should be reverted.

Polanyi and Zaenen (2006) introduced the concept of contextual valence shifter, which consists of negation, intensifier and diminisher. Contextual valence shifters have an impact of flipping the polarity, increasing or decreasing the degree to which a sentimental term is positive or negative.

But-clauses. The phrases like *"but", "with the exception of", "except that", "except for"* generally change the polarity of the part of the sentence following them. In order to handle these clauses, the opinion orientation of the text before and after these phrases should be set opposite to each other. However, there are situations when reversing the sentiment score of the second half of the sentence would be incorrect, for example, in the sentence *"Not only he is smart, but also very kind,"* the word "but" does not carry contrary meaning. These situations need to be considered separately.

Tokenisation into n-grams. Tokenisation is a process of creating a BOW from the text. The incoming string gets broken into comprising words and other elements, for example URL links. The common separator for identifying individual words is whitespace, however other symbols can also be used. Tokenisation of social-media data is considerably more difficult than tokenisation of the general text since it contains numerous emoticons, URL links, abbreviations that cannot be easily separated as whole entities. It is a general practice to combine accompanying words into phrases or n-grams: unigrams, bigrams, trigrams, etc. Unigrams are single words, while bigrams are collections of two neighbouring words in a text, and trigrams are collections of three neighbouring words. n-grams method can decrease bias, but may increase statistical sparseness. It has been shown that the use of n-grams can improve the quality of text classification (Raskutti et al., 2001, Zhang, 2003, Diederich et al., 2003), however there is no unique solution for the the size of n-gram. Caropreso et al. (2001) conducted an experiment of text categorization on the Reuters-21578 benchmark dataset. They reported that in general the use of bigrams helped to produce better results than the use of unigrams, however while using Rocchio classifier (Rocchio, 1971) the use of bigrams led to the decrease of classification quality in 28 out of 48 experiments. Tan et al. (2002) reported that use of bigrams on Yahoo-Science dataset allowed to improve the performance of text classification using Naive Bayes classifier from 65% to 70% break-even point, however, on Reuters-21578 dataset the increase of accuracy was not significant. Conversely, trigrams were reported to generate poor performances (Pak and Paroubek, 2010).

2.1.3 A Lexicon Based Approach

Lexicon-based approach calculates the sentiment of a given text from the polarity of the words or phrases in that text (Turney, 2002). For this method a lexicon (a dictionary) of words with assigned to them polarity is required. Examples of the existing lexicons include: Opinion Lexicon (Hu and Liu, 2004), SentiWordNet (Esuli and Sebastiani, 2006), AFINN Lexicon (Nielsen, 2011), LoughranMcDonald Lexicon, NRC-Hashtag (Mohammad et al., 2013), General Inquirer Lexicon[1] (Stone and Hunt, 1963).

The sentiment score of the text can be computed as the average of the polarities conveyed by each of the words in the text. The methodology for sentiment calculation can be described with the following steps:

- **Pre-processing.** The text undergoes pre-processing steps that were described

[1] http://www.wjh.harvard.edu/~inquirer/

in the previous section: POS tagging, stemming, stop-words removal, negation handling, tokenisations into n-grams. The outcome of the pre-processing is a set of tokens or a BOW.

- **Checking each token for its polarity in the lexicon.** Each word from the BOW is compared against the lexicon. If the word is found in the lexicon, the polarity w_i of that word is added to the sentiment score of the text. If the word is not found in the lexicon its polarity is considered to be equal to zero.

- **Calculating the sentiment score of the text.** After assigning polarity scores to all words comprising the text, the final sentiment score is calculated by dividing the sum of scores of words caring the sentiment by the number of such words:

$$\text{Score}_{\text{Avg}} = \frac{1}{m} \sum_{i=1}^{m} W_i. \tag{2.1}$$

  Sentiment score value ranges between -1 and 1, where 1 means a strong positive sentiment, -1 means a strong negative sentiment and 0 means that the text is neutral.

The quality of classification highly depends on the quality of the lexicon. Lexicons can be created using different techniques:

Manually construction of lexicons. The straightforward approach, but also the most time consuming, is to manually construct a lexicon and tag words in it as positive or negative. For example, Das and Chen (2001) constructed their lexicon by reading several thousands of messages and manually selecting words, that were carrying sentiment. They then used a discriminant function to identify words from a training dataset, which can be used for sentiment classifier purposes. The remained words were "expanded" to include all potential forms of each word into the final lexicon. Another example of hand-tagged lexicon is The Multi-Perspective-Question-Answering (MPQA) Opinion Corpus[2] constructed by Wiebe et al. (2005). MPQA is publicly available and consists of 8,222 subjective expressions along with their POS-tags, polarity classes and intensity.

[2] Available at `nrrc.mitre.org/NRRC/publications.htm`

Another resource is The SentiWordNet created by Esuli and Sebastiani (2006). SentiWordNet extracted words from WordNet[3] and gave them the probability of belonging to positive, negative or neutral classes, and subjectivity score. Ohana and Tierney demonstrated that SentiWordNet can be used as an important resource for sentiment calculation Ohana and Tierney (2009).

Constructing a lexicon from trained data. This approach belongs to the category of supervised methods, because a training dataset of labelled sentences is needed. With this method the sentences from the training dataset get tokenised and a BOW is created. The words are then filtered to exclude POS that do not carry sentiment, for example, prepositions. The prior polarity of words is calculated according to the occurrence of each word in positive and negative sentences. For example, if a word "success" is appearing more often in the sentences labelled as positive in the training dataset, the prior polarity of this word will be assigned a positive value.

Extending a small lexicon using bootstrapping techniques. Hatzivassiloglou and McKeown (1997) proposed to extend a small lexicon comprised of adjectives by adding new adjectives which were conjoined with the words from the original lexicon. The technique is based on the syntactic relationship between two adjectives conjoined with "and". It was established that "and" usually joins words with the same semantic orientation. Similarly, Hatzivassiloglou and McKeown (1997) and Kim and Hovy (2004) suggested to expand a small manually constructed lexicon with synonyms and antonyms obtained from NLP resources such as WordNet[4]. Other approaches include extracting polar sentences by using structural clues from HTML documents Kaji and Kitsuregawa (2007), recognising opinionated text based on the density of other clues in the text Wiebe and Wilson (2002). After the application of a bootstrapping technique it is important to conduct a manual inspection of newly added words to avoid errors.

2.1.4 A Machine Learning Based Approach

A Machine Learning approach for text classification is a supervised algorithm that analyses texts that were previously labelled as positive, negative or neutral; extracts features that model the differences between different classes, and infers a function,

[3] `http://wordnet.princeton.edu/`

[4] `https://wordnet.princeton.edu/`

that can be used for classifying new examples unseen before. In the simplified form, the text classification task can be described as follows: given a dataset of labelled data $T_{train} = \{(t_1, l_1), \ldots, (t_n, l_n)\}$, where each text t_i belongs to a dataset T and the label l_i is a pre-set class within the group of classes L, the goal is to build a learning algorithm that will receive as an input the training set T_{train} and will generate a model that will accurately classify unlabelled texts.

Some researches classified texts only as positive or negative (Pang et al., 2002), assuming that all the texts carry an opinion. Later, Wilson et al. (2005), Pak and Paroubek (2010) and Barbosa and Feng (2010) showed that short messages like tweets and blogs comments often just state facts. Therefore, incorporation of the neutral class into the classification process is necessary.

The process of machine learning text classification can be broken into the following steps:

1. **Data Pre-processing.** Before training the classifiers each text needs to be pre-processed and presented as an array of tokens. This step is performed according to the process described in section 2.1.2.

2. **Feature generation.** Features are text attributes that are useful for capturing patterns in data. The most popular features used in machine learning classification are the presence or the frequency of n-grams extracted during the pre-processing step. In the presence-based representation for each instance a binary vector is created in which "1" means the presence of a particular n-gram and "0" indicates its absence. In the frequency-based representation the number of occurrences of a particular n-gram is used instead of a binary indication of presence. In cases where text length varies greatly, it might be important to use term frequency (TF) and inverse term frequency (IDF) measures (Rajaraman and Ullman, 2011). However, in short messages like tweets words are unlikely to repeat within one instance, making the binary measure of presence as informative as the counts (Ikonomakis et al., 2005).

   Apart from the n-grams, additional features can be created to improve the overall quality of text classification. The most common features that are used for this purpose include:

- Number of words with positive/negative sentiment;
- Number of negations;
- Length of a message;
- Number of exclamation marks;
- Number of different parts-of-speech in a text (for example, number of nouns, adjectives, verbs);
- Number of comparative and superlative adjectives.

3. **Feature selection.** Since n-grams are the main features used in text classification, the dimensionality of the feature space grows proportionally to the size of the dataset. This dramatical growth of the feature space makes it, in most cases, computationally infeasible to calculate all the features of a sample. Feature selection is the process of identifying a subset of features that has the highest predictive power. This step is crucial for the classification process, since elimination of irrelevant and redundant features allows to reduce the size of feature space, increasing the speed of the algorithm, avoiding overfitting, as well as contributing to the improved quality of classification.

   There are three basic steps in feature selection process (Dash and Liu, 1997)

   (a) *Search procedure.* A process that generates a subset of features for evaluation. A procedure can start with no variables and add them one by one (forward selection) or with all variables and remove one at each step (backward selection), or features can be selected randomly (random selection).

   (b) *Evaluation procedure.* A process of calculating a score for a selected subset of features. The most common metrics for evaluation procedure are: Chi-squared, Information Gain, Odds Ratio, Probability Ratio, Document Frequency, Term Frequency. An extensive overview of search and evaluation methods is presented in (Ladha and Deepa, 2011a, Forman, 2003).

   (c) *Stopping criterion.* The process of feature selection can be stopped based on a: i) search procedure, if a predefined number of features was selected or predefined number of iterations was performed; ii) evaluation procedure, if the change of feature space does not produce a better subset or if optimal subset was found according to the value of evaluation function.

4. **Learning an Algorithm.** After feature generation and feature selection steps the text is represented in a form that can be used to train an algorithm. Even though many classifiers have been tested for sentiment analysis purposes, the choice of the best algorithm is still not easy since all methods have their advantages and disadvantages (see Marsland (2011) for more information on classifiers). The most popular learning algorithms for text classification are Support Vector Machines (SVMs) (Cortes and Vapnik, 1995, Vapnik, 1995), Naive Bayes (Narayanan et al., 2013), Decision Trees (Mitchell, 1996). Barbosa and Feng (2010) report better results for SVMs, while Pak and Paroubek (2010) obtained better results for Naive Bayes. In the work by Dumais et al. (1998) a decision tree classifier was shown to perform nearly as well as an SVM classifier.

2.1.5 Machine Learning Classifiers

2.1.5.1 Decision trees

Decision tree builds classification model in the form of a tree structure. In the process of building a tree, a training set gets split into smaller subsets, where each internal node represents a test on a feature, each branch denotes the outcome of a test and each leaf node represents a decision (class label). The top node in a tree which corresponds to the best predictor is called a root node. Fig. 2.1 depicts a simple tree for making a decision whether to play outside.

The core algorithm for building decision trees is ID3 proposed by Quinlan (1986). ID3 employs a top-down, greedy search through attributes. To choose the next attribute (feature) at each step, that splits the set of training instances into subsets with similar values, it uses Shannon Entropy measure (Shannon, 1948). Let S denote a set of training instances. Then, the entropy of set S is:

$$I(S) = -\sum_{x \in X} p(x) \log_2 p(x), \tag{2.2}$$

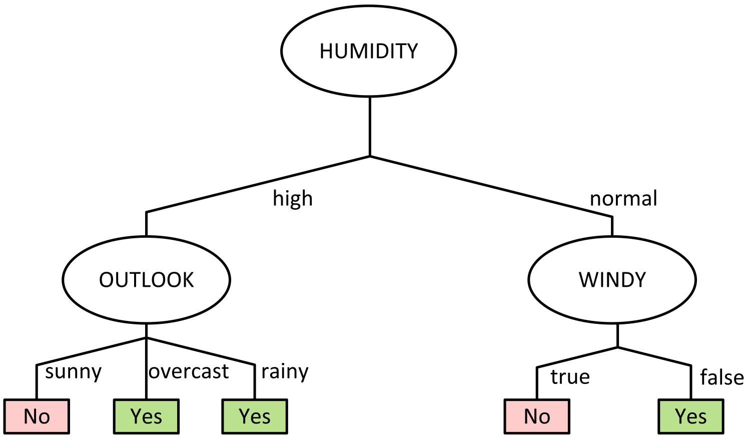


Figure 2.1: A decision tree to decide whether to play outside.

where X is a set of classes in S, $p(x)$ is the proportion of the number of elements in class x to the number of elements in set S.

The entropy of S conditional on attribute A is:

$$I(S|A) = -\sum_{v \in Values(A)} \frac{|S_v|}{|S|} I(S_v), \tag{2.3}$$

where $Values(A)$ are the subsets created from splitting set S by the attribute A, $|S|$ denotes the size of set S, $H(S_v)$ is the entropy of subset v.

Information Gain measures the decrease in entropy when the feature A is present vs. when it is absent:

$$G(S,A) = I(S) - I(S|A), \tag{2.4}$$

Constructing a decision tree is about finding attribute that returns the highest information gain. This process of choosing the most optimal attribute is repeated recursively until the stopping criteria is met. Possible stopping criteria are: number of cases in the node is less than some pre-specified limit, purity of the node is more than some pre-specified limit, depth of the node is more than some pre-specified limit, predictor values for all records are identical - in which no rule could be generated to split them.

Decision trees can be easily adapted to classifying textual data and have a number of useful qualities: they are relatively transparent, which makes them simple to understand; they give direct information about which features are important in making decisions, which is especially true near the top of the decision tree. However, decision trees also have a few disadvantages. One problem is that trees can be easily overfitted. The reason lies in the fact that each branch in the decision tree splits the training data, thus, the amount of training data available to train nodes located in the bottom of the tree, decreases. This problem can be addressed by using the tree pruning. The second weakness of the method is the fact that decision trees require features to be checked in a specific order. This limits the ability of an algorithm to exploit features that are relatively independent of one another.

2.1.5.2 Naive Bayes Classifier

Naive Bayes Classifier is frequently used for sentiment analysis purposes because of its simplicity and effectiveness. The basic concept of the Naive Bayes classifier is to determine a class (positive negative, neutral) to which a text belongs using probability theory. The algorithm uses a well-known Bayes Rule:

$$P(H \mid E) = \frac{P(H) * P(E|H)}{P(E)} \tag{2.5}$$

Where, P(H|E) – posterior probability of the hypothesis

$P(H)$ - prior probability of hypothesis

$P(E)$ - prior probability of Evidence

$P(E|H)$ conditional probability of Evidence given Hypothesis

In case of the sentiment analysis there will be three hypotheses – one for each sentiment class, and the one that has the highest probability will be selected as a class of the text. Following the Bayes rule, to find whether the text is positive, for example, it is necessary to compute the probability of a text given the positive class and the prior probability of positive class:

$$P(pos|E) = \frac{P(pos) * P(E|pos)}{P(E)} \tag{2.6}$$

$P(pos)$ and $P(E|pos)$ are obtained using the training set: $P(pos)$ will be counted as the number of instances labelled as positive divided by the total number of instances, $P(E|pos)$ can be approximated as a product of the probability of all words in a text belonging to a positive class:

$$P(E|pos) = \prod_{i=1}^{n} P(w_i|pos) \tag{2.7}$$

Where w_i is the word from a text, and the assumption is made that the occurence of word w_i does not depend on the presence of any other word w_j in the text. The only question is, what is the value of $P(E)$, however, since we are only interested in comparing the posterior probabilities for each class, denominator $P(E)$ can be ignored.

The potential problem with this approach appears if some word in the training set appears only in one class and does not appear in any other classes. In this case the classifier will always classify the text to that particular class. To avoid this effect Laplace smoothing technique could be applied.

2.1.5.3 Support Vector Machines Classifier

The Support Vector Machines were developed by Vapnik et al. [48] based on structural risk minimization principle. For the problem of separating the set of training vectors belonging to two classes, $\{(x_1,y_1),\ldots,(x_n,y_n)\}$, where y_i are either 1 or -1 indicating the class, the SVMs compute the hyperplane with maximum Euclidian distance to the closest training example. The distance is called margin and the hyperplane is called the maximum margin hyperplane. Mathematically, the points of the hyperplane should satisfy $wx_i + b = 0$ where w is a vector of weights normal to the plane and b is the bias. Classification rule is as follows:

$f(x) = sign\,(wx + b)$, given that all the instances must satisfy:

$$\begin{cases} wx_i - b \ \geq 1,\ for\ y_i = +1 \\ wx_i - b \leq 1,\ for\ y_i = -1 \end{cases} \tag{2.8}$$

The points that are the most close to the hyperplane are called *support vectors* and the margin is defined as twice their distance to the hyperplane $\frac{2}{\|w\|}$. In order to maximize the margin, the $\|w\|$ should be minimized. The maximum margin is depicted in Fig. 2.2.

When the data is non-linearly separable in R^N, it may be linearly separable in a higher-dimensional space R^M (where M ¿ N). In this context, a kernel function can be used to implicitly transform datasets to a higher-dimensional R^M space using

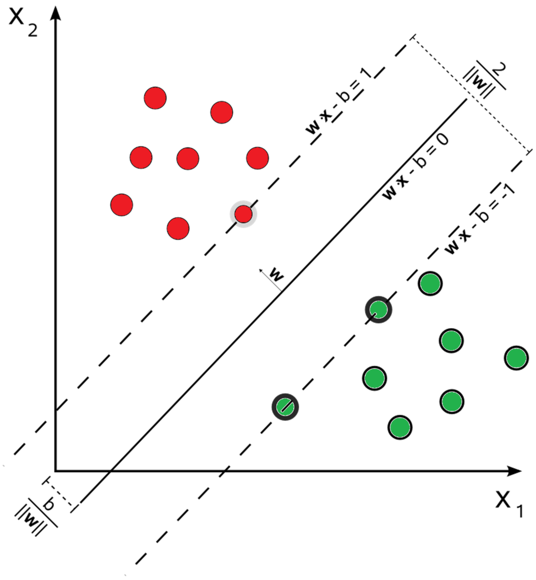


Figure 2.2: Maximum margin hyperplane.

no extra memory and with a minimal effect on computational time.

In a three-class sentiment classification scenario, there will be three pair-wise classifications: positive-negative, negative-neutral, positive-neutral. The method has proved to be very successful for the task of text categorization (Joachims, 1999, Dumais et al., 1998) since it can handle very well large feature spaces, however, it has low interpretability and is computationally expensive, because it involves calculations of discretisation, normalization and dot product operations.

2.1.6 Evaluation of Text Categorization

After the model is trained using a classifier it should be validated, typically, using a cross-validation technique, and tested on a hold-out dataset. There are several metrics defined in information retrieval for measuring the effectiveness of classification: accuracy, error rate, precision, recall and F-score. A typical confusion table for classification problem is given in Table 2.1, where TN means true negatives, TP means true positives, FP stands for false positives and FN corresponds to false negatives.

Table 2.1: Confusion Matrix

True class	**Predicted class**	
	YES	NO
YES	TP	FN
NO	FP	TN

- **Accuracy.** Kotsiantis (2007) defined accuracy as "the fraction of the number of correct predictions over the total number of predictions":

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{2.9}$$

  Accuracy allows to measure the number of correctly predicted instances and does not give information about the number of incorrect predictions. For this purposes the Error Rate measure can be used.

- **Error rate.** Error rate measures the number of incorrectly predicted instances against the total number of predictions. Computationally, error rate is (1 - Accuracy) on the training and test examples. The formula for calculating the error rate is presented in Eq. 2.10.

$$\text{Error rate} = \frac{FP + FN}{TP + FP + FN + TN} \tag{2.10}$$

- **Precision.** Precision shows the proportion of how many instances the classifier got right to the total number of true positive and true negative examples. Precision shows the exactness of the classifier with respect to each class:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2.11}$$

- **Recall.** Recall represents the proportion of the number of texts, which were assigned a positive sentiment score, to the total number of true positives and false negatives. Recall shows the completeness of the classifier with respect to each class. It can identify the class, which is the most difficult for a classifier to predict:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2.12}$$

  There is a discussion about the advantages of using accuracy and error rate over precision and recall. Manning and Schütze (1999) argue, that for unbalanced datasets precision and recall can be better metrics for measuring classifiers' performance. There can be the trade-off between precision and recall, since one can be increased at the expense of the other. For example, in some extreme cases the recall can be up to 100 %, but precision can be very low. It would not be possible to correctly evaluate the classifier in this scenario. In these cases the F1-score metric can be used, which combines the values of precision and recall.

- **F1-score.** Rijsbergen (1979) defined the F1-score as the harmonic mean of precision and recall:

$$\mathrm{F-Score} = \frac{2 * \mathrm{Precision} * \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{2.13}$$

Depending on the nature of the task, one may use accuracy, precision, recall or F1-score as a metric or some mixture of them.

2.2 Time Series Analysis

2.2.1 Definition of Time Series

Time Series is a collection of observations of a variable obtained through repeated measurements at equally spaced time intervals. Time series are important, because they allow to analyse phenomena that changes over time. The classical examples of time series are stock prices, meteorological data like temperature or rainfall, the value of retail sales each month of the year. Time series analysis finds applications in many areas of science and engineering, for example: economic forecasting, consumer demand forecasting, inventory study, stock market predictions.

Two main objectives of analysing time series are:

1. Understanding the nature of the phenomena.

2. Developing a quantitative forecasting models that can be used to forecast future data as a function of past data.

2.2.2 Definition of Stationarity

The important step before doing time series analysis is a stationarity check. Challis and Kitney (1991) define stationarity as "a quality of a process in which the statistical parameters (mean and standard deviation) of the process do not change with time". There are two types of stationarity: strict and weak. As per Tsay (2005):

1. ***Strictly stationary*** time series $\{r_t\}$ is the time series for which joint distribution of $(r_{t_1}, \dots, r_{t_k})$ is identical to that of $(r_{t_{1+t}}, \dots, r_{t_{k+t}})$ for all t, where k is an arbitrary positive integer and $(t_1, \dots, t_k)$ is a collection of k positive integers. In other words, for the strictly stationarity process the joint distribution of $(r_{t_1}, \dots, t_{t_k})$ is invariant under time shift.

2. ***Weakly stationary*** is the time series $\{r_t\}$ for which both mean of r_t and the covariance between r_t and r_{t-l} are time-invariant, where l is an arbitrary integer. In practice, if one have observed T data points $\{r_t \mid t = 1, \dots, T\}$ of the weakly stationary process, the time plot of the data would show that the T values fluctuate with constant variation around a fixed level.

It is worth mentioning, that stationarity is a relative term. It is difficult to find a strongly stationary process in real life, it can only be described mathematically. Sometimes one can find weakly stationary time series and assume that they are "close enough" to the strongly stationarity processes to be treated as such. When the process is stationary in real life, in order to be seen as stationary mathematically, the data should be collected for a very long period of time when compared to the total length of the data. Taking this into account, if the data for a stationary process was collected for only a short period of time, then the process will appear to be non-stationary.

When analysing time series, it is very important to have the property of stationarity, because it allows making many simplifying assumptions and applying statistical tools to the data. In contrast, non-stationary data is unpredictable and cannot be used for forecasting and modeling. For example, if the correlation is found between two time series which are non-stationary, one may conclude that there is a relationship between the variables, however, this relationship not necessarily exists in real life. This phenomena is called "spurious regression" and it has been extensively studied by Yule (1926) and Granger and Newbold (1974) for economic data. For more information on spurious regression, see Ventosa-Santaularia (2009).

2.2.3 Detection of Non-Stationarity

In order to make reliable conclusions about time series one should transform non-stationary data into the stationary data. Before applying the transformation, however, one should understand which type of non-stationarity he is dealing with: trend, seasonality or their combination. Any time series can be decomposed into the three components:

$$X_t = m_t + s_t + \epsilon_t \tag{2.14}$$

where m_t is a trend component, which changes slowly over time; s_t is a seasonal component and can be described mathematically through a function with a known period; and ϵ_t is a random noise component.

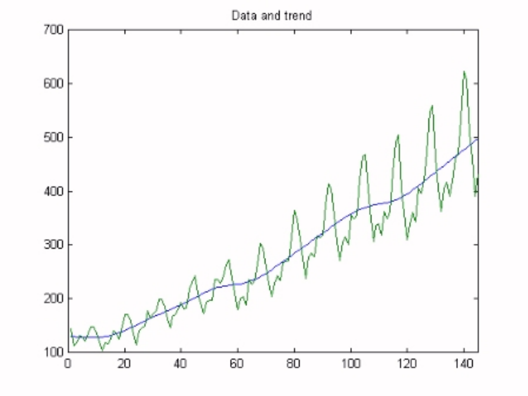


Figure 2.3: Time series with a trend component (blue line).

Time series which have trends, seasonality or the combination of both are not stationary. This is because both, trend and seasonality will have an impact on the values of time series.

In order to identify if a trend is present in the time-series it is often enough to analyse a simple sequence plot. Fig. 2.3 clearly shows the presence of an upward trend. Detecting seasonality can be sometimes difficult using just the sequence plot. One can try using the following techniques: assessment of autocorrelation function (ACF) (Box and Jenkins, 1976); Kwiatkowski-Phillips-Schmidt-Shin Test (Kwiatkowski et al., 1991); Augmented Dickey-Fuller Test (Dickey and Fuller, 1979); Phillips-Perron Test Phillips and Perron (1988).

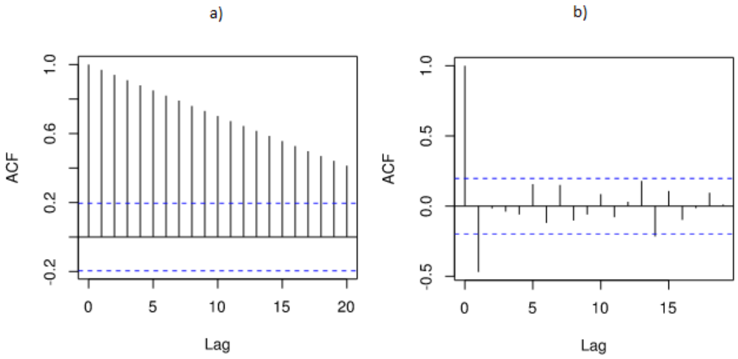


Figure 2.4: ACF correlograms for a) non-stationary time-series; b) stationary time series.

2.2.3.1 Evaluation of ACF correlogram.

Autocorrelation plot allows to check the randomness of data by computing correlations at different time lags. If the data is random, the ACF must drop to levels below significance with exponential or faster rate, while the ACF of non-stationary data declines slower than exponentially over a prolonged period of time. Fig. 2.4 shows examples of a) ACF for non-stationary time-series; b) ACF for stationary time series.

2.2.3.2 Augmented Dickey-Fuller Test.

A stationary variable tends to return to a fixed mean after being disturbed by a shock (mean-reverting). This tendency to revert back to the mean is the basis for the most popular test of stationarity, the Dickey-Fuller test (DF) (Dickey and Fuller, 1979). The test has several variants depending on whether a non-zero constant and/or a deterministic trend is included. Suppose a variable y is characterised by the basic AR(1) process (Parker, 2017):

$$y_t = \rho y_{t-1} + u_t, \tag{2.15}$$

where u_t is assumed to be stationary. If one subtracts y_{t-1} from both sides:

$$\Delta y_t = (\rho - 1) y_{t-1} + u_t = \gamma y_{t-1} + u_t, \tag{2.16}$$

with $\gamma \equiv \rho - 1$. The null hypothesis H_0 that variable y is non-stationary is equivalent to $\gamma = 0$ with the alternative $\gamma < 0$ (Parker, 2017).

The DF test is valid if u_t in Eq. 2.15 is white noise, but in real life the error terms in time series are often correlated. Augmented Dickey-Fuller Test (ADF) allows to eliminate the autocorrelation in u by including lagged values of Δy_t. Mathematically, the ADF test with a constant (drift), a linear trend and p lags can be described as follows:

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \phi_1 \Delta y_{t-1} + \ldots + \phi_p \Delta y_{t-p} + \epsilon_t, \tag{2.17}$$

where α is a constant, β represents the coefficient of a deterministic trend, ϵ_t is en error term $\sim iid(0,{\sigma_\epsilon}^2)$ by construction, Δy_t denotes lagged first-differenced series, $\Delta y = y_t - y_{t-1}$, and p is the number of lags to include in the regression, which has a strong impact on the outcome of the unit root tests. Imposing the constraints $\alpha = 0$ and $\beta = 0$ corresponds to modelling a random walk and using the constraint $\beta = 0$ corresponds to modelling a random walk with a drift.

The role of the ADF hypothesis test is to consider the null hypothesis that $\gamma = 0$ (the presence of a unit root), which would indicate that there is no tendency for high values of y in $t-1$ to be reversed in t, thus, the process is not mean-reverting (is non-stationary). If the calculated ADF test statistic is less than the negative critical value at our desired level of significance, then one should reject the null hypothesis and conclude that the series is stationary. If the test statistic is positive or less negative than the critical value, then one cannot reject the hypothesis that y is non-stationary. It is important to note that depending on whether a drift and/or a deterministic trend is present, different sets of critical values should be used for the test.

2.2.3.3 Phillips-Perron Test.

Phillips-Perron test (Phillips and Perron, 1988) builds on the Dickey-Fuller test and has the same null hypothesis. However, whilst the augmented Dickey-Fuller test addresses the issue of higher order of autocorrelation by introducing lags of Δy_t as regressors in the test equation, the PP test corrects for any serial correlation and heteroskedasticity in the errors non-parametrically by modifying the Dickey-Fuller test statistics, so that no additional lags of the dependent variable are needed.

A great advantage of Philips-Perron test is that it assumes no functional form of the error (non-parametric nature), which makes it applicable to a wide range of problems. However, the disadvantage of this test is that it is based on asymptotic theory and, therefore, it works well only in large samples. It has been demonstrated that the Phillips-Perron test performs worse in small samples than the augmented Dickey-Fuller test (Davidson and MacKinnon, 2004).

2.2.3.4 Kwiatkowski-Phillips-Schmidt-Shin Test.

KPSS test (Kwiatkowski et al., 1991), in which the null hypothesis is that of stationarity, is intended to complement unit root tests. KPSS test can be described using the following form :

$$y_t = \delta t + \varsigma_t + \epsilon_t, \tag{2.18}$$

where δt is a deterministic trend, ϵ_t is a stationary error and

$$\varsigma_t = \varsigma_{t-1} + u_t, \text{with } u_t \sim iid(0, {\sigma_u}^2) \tag{2.19}$$

such that it follows a random walk. A test of ${\sigma_u}^2 = 0$ is a test for stationarity. See Maddala and Kim (1998), Virmani (2004) for further details.

The authors of KPSS test derived one-sided LM statistics for the test. If the LM statistic is greater than the critical value, then the null hypothesis is rejected, thus the series is non-stationary.

Opposite to PP test, KPSS is a more powerful in small samples (Maddala and Kim, 1998). However, a major disadvantage for the KPSS test is that it has a high rate of Type I errors (rejecting the null hypothesis too often). It is suggested that in order to deal with the Type I errors, KPSS tests should be combined with unit root tests, such as ADF and PP.

2.2.4 Elimination of Non-Stationarity

Transformations such as logarithms can help to stabilise the variance of a time series. Differencing can help stabilise the mean of a time series by removing changes in the level of a time series, and so eliminating trend and seasonality. Differencing is the transformation of the data x_t to a new data d_t, where the values d_t are the differences between sequential data points of X_t:

$$d_t^{'} = x_t - x_{t-1} \tag{2.20}$$

Sometimes even after the differencing operation the data still remains non-stationary. The second-order differencing may help to obtain the stationary data:

$$d_t^{''} = d_t^{'} - d_{t-1}^{'} = (x_t - x_{t-1}) - (x_{t-1} - x_{t-2}) = x_t - 2x_{t-1} + x_{t-2} \tag{2.21}$$

A seasonal difference is the difference between the present time series and the corresponding values from the previous year. So, if m is the number of seasons:

$$d_t^{'} = x_t - x_{t-m} \tag{2.22}$$

The resulted time-series are called lag-m differences. If the data after the seasonal difference operator appears to be white noise, then the appropriate model for the original data will be:

$$x_t = x_{t-m} + \epsilon_t \tag{2.23}$$

Sometimes one should apply seasonal difference and first difference in order to obtain stationary data. In case when strong seasonality is present the seasonal differencing should be done prior to the first differencing, because the residuals may become stationary after the first step and there may be no need for the first differencing.

2.2.5 Definition of Cointegration

As it has been discussed in the previous section, regressing two non-stationary variables results in spurious regression. One extremely important exception to this rule is the case of cointegration (first discussed by Granger (1981), Granger and Weiss (1983), Engle and Granger (1987)). If the series y_t and x_t are both non-stationary, but are cointegrated, spurious regression no longer arises. Cointegration means that there is a linear combination of y_t and x_t that is stationary I(0). If two or more I(1) variables are cointegrated, they must obey an equilibrium relationship in the long-run, although they may diverge substantially from that equilibrium in the short run.

2.2.6 Cointegration Testing

To explore whether sales and social-media factors share an underlying cointegrated relationship, one can use Johansen test (Johansen, 1988, 1991). Johansen test allows to determine if three or more time series are cointegrated. It starts with the vector auto-regression (VAR) of order p:

$$y_t = A_1 y_{t-1} + \ldots + A_p y_{t-p} + u_t, \tag{2.24}$$

where y_t is an $n \times 1$ vector of variables that are integrated of order one I(1), A_i are the coefficient matrices for each lag and u_t is a multivariate Gaussian noise term with mean zero. After subtracting y_{t-1} from both sides (differencing) and rearranging the terms this VAR can be re-written as:

$$\Delta y_t = \Pi y_{t-1} + \sum_{i=1}^{p-1} \Gamma_i \Delta y_{t-i} + u_t, \tag{2.25}$$

where

$$\Pi = \sum_{i=1}^{p} A_i - I \text{ and } \Gamma_i = -\sum_{j=i+1}^{p} A_j. \tag{2.26}$$

If Π is equal to zero this means that there is no cointegration. If Π has reduced rank, $r < n$ and is not equal to zero, it is the case of cointegration, in which Π can be written as $\Pi = \alpha\beta'$, where α and β are $n \times r$ matrices with rank r and $\alpha'\beta$ is stationary. r corresponds to the number of cointegrating relationships, α are the adjustment parameters in the Vector Error Correction model and each column of β is a cointegrating vector.

Johansen proposed two different likelihood ratio tests to determine the rank r: the trace test and maximum eigenvalue test:

$$J_{\text{trace}} = -T \sum_{i=r+1}^{n} \ln(1 - \hat{\lambda}_i) \tag{2.27}$$

$$J_{\max} = -T \ln(1 - \hat{\lambda}_{r+1}) \tag{2.28}$$

Where T is the sample size and $\hat{\lambda}_i$ is the i:th largest canonical correlation of Δy. The trace method tests the null hypothesis of r cointegrating vectors against the alternative hypothesis of n cointegrating vectors. The maximum eigenvalue test, on the other hand, tests the null hypothesis of r cointegrating vectors against the alternative hypothesis of $r + 1$ cointegrating vectors. For more information on Johansen test refer to (Hjalmarsson and Osterholm, 2007).

2.3 Events Detection and Clustering

2.3.1 Event Study: Related Work

Event study was introduced for the first time by Dolley (1933), and was used to measure the effects of economic events on the value of companies. In general, events can be defined as real-world unique occurrences that unfold over space at some point in time (Atefeh and Khreich, 2015, Allan et al., 1998). Events are present in all natural and social systems: diseases and epidemics, economic draw-downs, financial stock markets crashes, retail sales spikes, political protests, etc. Identification of precursors and warning signals of such events can have tremendous social and economic impacts (Piovani et al., 2015), from preventing the spread of diseases, mitigating riot threats to modifying business strategies and performing economic structural adjustments.

In the recent years there has been a growing interest in event detection using publicly available data from online social media sources. People use blogs, Facebook, Twitter, Instagram to discuss their ideas, express opinions, share information. Hence, Twitter, for example, became a reliable source for detecting breaking news (Li et al., 2016, Petrovic et al., 2013, Petrović et al., 2012, Li et al., 2012, Weng and Lee, 2011, Phuvipadawat and Murata, 2010). Information from social media can also be extremely helpful for gaining greater insights into crowd behaviours or collecting relevant information about real-world events as they unfold. The recent work by Alsaedi et al. (2017) showed that Twitter posts can be used to detect riot events in real time. The authors demonstrated that their framework exposed riots faster than information about them was reported to the London Metropolitan Police Service. Previously, Alsaedi and Burnap (2015) were able to identify real-world events in Arabic using Twitter. Similarly, Kallus (2014) accurately predicted significant protests by analysing a massive dataset from various open-content online sources, such as news, government publications, blogs, social-media, while Alanyali et al. (2016) used Flickr data to identify protest outbreaks. Twitter has also been used to provide timely information about disasters allowing to speed up emergency response (Yin et al., 2015, Imran et al., 2014, Yin et al., 2014, Sakaki et al., 2010, Vieweg et al., 2010). Preis and Moat (2014) showed that search data from Google Trends can help to estimate the current number of influenza infections; Radinsky and Horvitz (2013) were able to predict cholera disease outbreaks as the results of natural disasters by analysing news stories.

As demonstrated above, most existing approaches to event detection focus on global or large-scale event detection (breaking news, massive protests, natural disasters) (Alsaedi et al., 2017). However, social media information can also be useful for detecting smaller-scale events. One of the domains that can leverage information from social media is a business domain. Similarly to how active protesters use Twitter and other social networks to reinforce their ideas and call other people to action, people talk about their intent to buy specific products or share their opinion about already purchased goods, triggering other potential customers to buy (or not) a specific product. Identifying spikes in online sentiments about brands and products, as well detecting purchase intentions by text analysis, might provide an indication of the future customer demand. These social media signals can be used by business managers to foresee potential spikes in sales, and to make more informed short-term and long-term decisions. While a lot of research work has been dedicated to sales forecasting, to my best knowledge, there are no extensive studies that utilise social media with the goal of predicting spikes (events) in sales of a company. The only identified research that mentions evaluation of Twitter events in relation to sales events is the study by Dijkman et al. (2015). Dijkman et al. (2015) found a significant correlation between spikes in volume of positive tweets and sales for the next few weeks. However, the findings are based on a very small Twitter dataset (12,780 tweets) and are limited to one company, thus, cannot be generalised. The studies (Ranco et al., 2015, Sprenger et al., 2014) did not analyse companies' sales, but found correlations between Twitter events and the abnormal market returns. Sprenger et al. (2014) showed that different categories of events had a different effect on the market, for example, events regarding Earnings and Mergers and Acquisitions contained information that moved the market, while Product Development and Joint Ventures event types had no effect on the markets. Ranco et al. (2015) differentiated between Earnings Announcements(EA) events and non-EA events and showed that while EA events had a stronger impact on the returns, non-EA events also had a significant relationship with the market. Both, Sprenger et al. (2014) and Ranco et al. (2015) provided empirical confirmation that Twitter sentiment is a stronger signal than Twitter volume.

Processing social media messages in search of relevant information is a challenge and involves solving multiple tasks, such as data collection, extracting, filtering, feature generation, feature selection, classification, ranking, aggregation and summarising. A vast amount of methods can be used to solve these tasks, including data mining, natural language processing, information retrieval, and machine learning. A detailed survey of the state of the art methods is presented in the work by Imran et al. (2015). In section 4 of the paper the authors cover methods for programmatic extraction of data from social media and its pre-processing, section 5 describes methods for events detection and section 6 discusses supervised and unsupervised methods for events classification and existing techniques for information summarisation. The authors also discuss the advantages and disadvantages of each method. In their recent work Liu et al. (2016a) reviewed research efforts related to existing studies on event detection in social multimedia data, which include not only textual information, but also images and video. An extensive survey of techniques specifically for Twitter, classified by event type, detection task, and detection method, was carried out by Atefeh and Khreich (2015). During the recent years many researchers created frameworks for events detection and clustering utilising the mix of methods (Shao et al., 2017, Yang et al., 2017, Li et al., 2016, Liu et al., 2016b, Qian et al., 2015, Corley et al., 2013, Reuter and Cimiano, 2012, Yang et al., 2012).

The analysis of the techniques, described in the surveys mentioned above, revealed that the differentiation between different types of social media events in existing literature is based mainly on temporal and textual information (Abdelhaq et al., 2017, Qian et al., 2015, Thapen et al., 2016, Corley et al., 2013, Yang et al., 2012, Werner and Murray, 2004), temporal and spacial information (Zhou et al., 2016, Zhao et al., 2016, Zhang et al., 2016, Krumm and Horvitz, 2015, Schulz et al., 2015, Dong et al., 2015), or the combination of temporal, textual and spacial features (Alsaedi et al., 2017, Cheng and Wicks, 2014, Zhou and Chen, 2014, Kallus, 2014).

While time, location, topic and content of the discussions are, undoubtedly, important, there is space to explore other factors that might be relevant in the context of social networks, for example, the dynamic of how information spreads through the social network. Some research efforts were done in this direction by including the network structure (Gao et al., 2017, Shao et al., 2017, Wenbin et al., 2017, Hu et al., 2017, Shao et al., 2017, Prangnawarat et al., 2015, Rozenshtein et al., 2014). Matsubara et al. (2017), Sano et al. (2013), Yang and Leskovec (2011), Crane and Sornette (2008), Sornette et al. (2004) studied the growth and relaxation signatures of spikes in order to understand the process of spikes formation and dissipation. Crane and Sornette (2008), Sornette et al. (2004) showed that dynamic responses of a social network to large social events is different depending on whether the event is endogenous or exogenous. For example, a catastrophic Asian Tsunami in December 2004 (exogenous nature) caused an immediate response in social networks, characterised by a sudden peak and relatively quick relaxation. Differently, spike of search activity related to the release of Hurry porter movie (endogenous word-of-mouth activity) had a relatively slow growth preceding the release and an almost symmetric decay of interest (Crane and Sornette, 2008).

Accounting for endogenous/exogenous dynamics can be relevant when analysing social media activities in relation to brands. Most business today are subjected to external stimulations, for example, marketing campaigns, promotions, IPOs. A discussion on Twitter will evolve differently through time depending on whether it was initiated by the brand through a marketing campaign or became a result of a word-of-mouth information sharing. A post-event effect on sales might also be different for the two scenarios. Differentiation between different events' dynamics by analysing spikes signatures presents an opportunity and will be explored in this research work on the case-study of Twitter events related to brands. To my best knowledge, no research has previously been done in analysing social media spikes' shapes for sales prediction.

In the following sections I will describe techniques for events detection and clustering relevant to this research work. Methods for event detection and clustering based on the textual, spacial and graphs information are outside the scope of this thesis. An interested reader might refer to the surverys by Imran et al. (2015), Liu et al. (2016a), Atefeh and Khreich (2015) in search of such information.

2.3.2 Event Detection Techniques

Events detection is a challenging task since the signal is often corrupted with noise. A general approach used in event study field is identifying the spikes (outliers) in data as single data points (Zhu and Guo, 2017, Liu et al., 2016a, Krumm and Horvitz, 2015) and defining an event window that will represent the duration of the event (Konchitchki and OLeary, 2011, Bessembinder et al., 2009, Brown and Warner, 1985).

An outlier is defined as an observation that is considerably different from the remainder of the dataset. The most popular outlier detection methods are:

1. **Extreme Studentized Deviation (ESD).** (Rosner, 1983). For a sample $x_N = \{x_i\}_{i=1}^N$ ESD classifies any point more than t standard deviations away from the mean to be an outlier, where the threshold value t is most commonly taken to be 3. In other words, x is identified as an outlier if:

$$|x - \overline{X}| \geq t\sigma \tag{2.29}$$

where $\overline{X}$ is the mean and σ is the estimated standard deviation of the data sequence.

$$\overline{X} = \frac{\sum_{i=1}^N x_i}{N},\ \sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{X})^2}{N-1}} \tag{2.30}$$

Motivation for the threshold choice $t = 3$ comes from the fact that for normally-distributed data, the probability of observing a value more than three standard deviations from the mean is only about 0.3%.

2. **Hampel identifier.** Hampel (1971, 1974). For this outlier detection method the mean is replaced with the median of the residuals and the standard deviation is replaced with the median absolute deviation estimate (MAD). MAD is a robust measure of the variability of univariate data. To compute

MAD, one calculates the median of the absolute deviations of each historical value from the data's median.

$$\mathrm{MAD}(X) = \mathrm{median}(|X_1 - \mathrm{median}(X)|, \ldots, |X_N - \mathrm{median}(X)|), \tag{2.31}$$

where N is the number of observations. x is identified as an outlier if:

$$|x - \mathrm{median}(X)| \geq g(N, \alpha_N)\mathrm{MAD}(X) \tag{2.32}$$

Where g is a function related to the number of data points and a specified type I error (Liu et al., 2004, Davies and Gather, 1993).

3. **Median and InterQuartile Range (IQR).** Tukey (1977). For this outlier detection method, one calculates the 25th percentile and the 75th percentile of the data. The difference between the 25th and 75th percentile is the interquartile deviation IQR. The historical value x is classified as an outlier if it is outside of the closed range:

$$[Q_1 - K * IQR;\ Q_3 + K * IQR] \tag{2.33}$$

where $IQR = Q_3 - Q_1$, Q_1 and Q_3 are the 25th and the 75th percentiles respectively, and K is often selected equal to 1.5.

2.3.3 Events Clustering Techniques

The process of events clustering based on their shapes is equivalent to a problem of time series clustering. Many clustering algorithms have been developed for time series data which can be grouped into two main categories:

1. *Feature-based.* These methods summarise or transform raw data into feature-set by using feature extraction techniques or parametric models. The selection of the appropriate features is the hardest part of the process. The typical features used in time series clustering are: time series frequency components extracted using discrete wavelet decomposition (Mallat, 1989),

spikes characteristics (mean spike height and width, mean inter-spike interval, standard deviation of spike heights and width, etc.) (Nowotny et al., 2013). Constructed features obtained from each individual series are then used in arbitrary clustering algorithms, including a self-organizing map or hierarchical clustering algorithm. The feature-based approach significantly reduces the dimensionality of data, however, it leads to data loss.

2. *Distance-based (shape-based, raw-data-based).* These methods are applied over raw time series without any transformation performed on data prior to clustering. To compare time sequences and find the similarities between them, the distance between data points of different time series is calculated, therefore, the outcome of the clustering critically depends on the choice of a distance/similarity metric. The most common measures are the Euclidean Distance (ED) and Dynamic Time Warping (DTW).

In this work I focus on the distance-based approach for two reasons: 1) since events are represented by short sub-sequences of the original time series, data remained after feature extraction might be not sufficient for accurate clustering using the feature-based approach; 2) conclusions can be easier generalised for other applications.

For more information on time series clustering refer to Aghabozorgi et al. (2015) that provide an extensive decade review of the techniques.

2.3.3.1 Distance Measures

As mentioned in the survey by Wang et al. (2006), in the last decade multiple similarity metrics have been proposed: autocorrelation (Wang and Wang, 2000), cepstrum (Kalpakis et al., 2001), piecewise normalization (Indyk et al., 2000), cosine wavelets (Huhtala et al., 1999), piecewise probabilistic measures (Keogh and Smyth, 1997). However, when tested on the same datasets, the best performing metrics are Euclidean Distance (Keogh and Kasetty, 2003) and Dynamic Time Warping (Berndt and Clifford, 1994, Ratanamahatana and Keogh, 2005) for time series of varying length. In the current work, I focus on these two general-purpose popular distances. For more information on other distance metrics for time series clustering refer to Liao (2005).

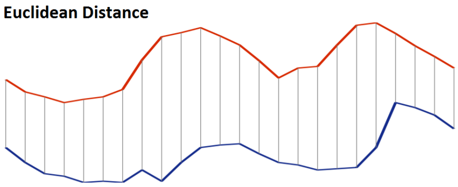


Figure 2.5: Linear mapping between two time series using Euclidean Distance

- *Euclidean Distance (ED)* ED calculates the similarity between two sequences of the same length by summing the ordered point-to-point distance between them (Eq. 2.34).

  $$d(T,S) = \sqrt{\sum_{i=1}^{n} (T_i - S_i)^2}, \tag{2.34}$$

  where T and S are time series of length n.

  The advantage of this metric is that it can be computed in linear time to the length of time series O(n). However, there are few limitations of ED: a) it requires the time series to be of the same length; b) it is sensitive to distortion in time (distortion in time appears in speech recognition, for example, where the speed of speech is not constant (Sakoe and Chiba, 1978)).

- *Dynamic Time Warping Distance (DTW)*

  While ED performs a linear map between points (see Fig. 2.5), DTW allows non-linear mapping (see Fig. 2.6) (Berndt and Clifford, 1994, Keogh and Ratanamahatana, 2005). It is a more robust measure than ED because it allows time shifting and thus matches similar shapes even if they are out-of-phase (Berndt and Clifford, 1994, Ratanamahatana and Keogh, 2005).

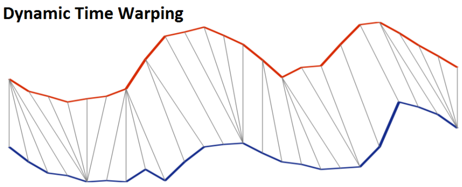


Figure 2.6: Non-linear mapping between two time series using Dynamic Time Warping

Given two time series $T = \{t_1, t_2, ..., t_n\}$ and $S = \{s_1, s_2, ..., s_m\}$ of length n and m respectively, a distance matrix $n * m$ is constructed where each element

(i^{th}, j^{th}) contains the Euclidean distance $d(t_i, s_j)$ between the two points t_i and s_j. The objective of DTW is to find the warping path $W = \{w_1, w_2, ..., w_k\}$ of contiguous elements in the distance matrix that minimizes the following function:

$$\text{DTW}(T, S) = \min \left(\sqrt{\sum_{k=1}^{K} w_k} \right), \tag{2.35}$$

where the k^{th} element of W is defined as $w_k = (i, j)_k$.

The main drawback of this measure is that finding the warping path W is computationally very expensive (complexity is O(n*m)). Knowing that the minimal path falls around the diagonal of the matrix, some researches tried to exploit this fact in order to speed up the DTW calculations (Xi et al., 2006).

2.3.3.2 Clustering Algorithms

In this section I review the major and most widely used techniques for shape-based time series clustering. The two most popular approaches are k-means clustering and hierarchical clustering.

- *K-means clustering*

  The basic intuition behind k-Means is grouping together similar objects and, using an iterative refinement technique, minimising the objective function. The algorithm works as follows (Kotsakos et al., 2013):

  1. Decide on the value of *k* which indicates the number of clusters;
  2. Initialise *k* cluster centres (if necessary, randomly)
  3. Assign objects to the most similar clusters using some distance function, for example, ED or DTW.
  4. Recalculate *k* cluster centres by averaging the objects of each cluster
  5. Repeat steps 3 and 4 until the centres of clusters don't change. The objective function, which is the sum of squared errors between cluster' centres and clusters' objects has been minimised.

  The choices of the initial centres are critical to the quality of results as K-Means is a hill-climbing algorithm and might converge on a local and not necessarily global optimum. The main disadvantage of the algorithm is that it requires to specify the number of clusters k a priori. Nevertheless, k-means is very poplar for clustering time series objects due to its fast running time. The time complexity of the algorithm is $O(kNrD)$, where N is the number of objects, r is the number of iterations until convergence and D is the dimensionality of data points.

- *Hierarchical clustering*

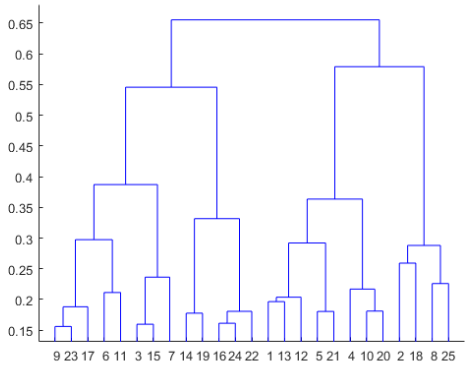


Figure 2.7: Hierarchical clustering dendrogram

Hierarchical clustering is one of the most widely used clustering approaches due to its great visualization power (see Fig. 2.7). The idea behind the algorithm is to create a nested hierarchy of similar groups of objects in which clusters in the higher level of hierarchy are created by merging the clusters from the next lower level (Hastie et al., 2009). In order to decide how the merging is performed, a (dis)similarity measure between groups should be specified (for example, single linkage, complete linkage, average linkage), in addition to the chosen pairwise distance measure (for example, ED or DTW).

There are two types of hierarchical clustering, agglomerative and divisive. In agglomerative procedures, every data objects is assign to its own cluster as the start, then objects are grouped together sequentially based on the similarity measure until the desired number of clusters is reached or until all objects are contained in a single cluster. Divisive procedure is the inverse of the agglomerative clustering, where the process starts with all data objects being in one cluster and continues by dividing them into smaller groups until each member is in a singleton.

The big advantage of the hierarchical clustering algorithm is its generality: the number of clusters k does not need to be specified a priory; also, the clustering process is deterministic, producing the same result for a chosen set of (dis)similarity measures. However, the application of hierarchical clustering is limited to only small datasets, since its time complexity is $O(N^2)$, where N is the total number of objects.

2.4 Time Series Forecasting

2.4.1 Application of Social Media for Forecasting: Related Work

Forecasting future sales is one of the key activities for any business. Indeed, accurate forecasts can help management to plan activities in marketing, production, purchasing, finance and accounting (Mentzer and Bienstock, 1998).

The majority of literature on retail sales forecasting is addressing univariate time series based on trend, seasonality and autocorrelation structures (Chu and Zhang, 2003, Alon et al., 2001), however, in the world of digital marketing and social networks, market trends and sentiments change very rapidly, making it questionable whether historical data remains the sole relevant variable for predicting current dynamics of consumer demand (Gur Ali and Pinar, 2016). Today people are used to express their opinions about products and services online and their decisions about purchasing are being greatly affected by other people's feedback. In this context, the number of Facebook likes has become a social reference system that represents the aggregated opinion of consumers (Lee et al., 2016), while Twitter has become the key source for posting information about promotions, and Google Trends became a reliable resource to can quickly determine what topics are grabbing people's interest. Utilising data from online sources might provide additional information about changing consumer interest and greatly improve the accuracy of forecasting models.

A large amount of studies have demonstrated the role of Twitter, Facebook, blogs activities, reviews and Google searches for predicting different social phenomena. For example, Twitter has been shown useful for estimating influenza rates and alcohol sales volume with high accuracy (Culotta, 2013), for predicting unemployment rates (Antenucci et al., 2014) and election results (Bermingham and Smeaton, 2011, Tumasjan et al., 2010, Kim and Hovy, 2007). The study by Dickinson and Hu (2015) demonstrated the presence of significant correlations between Twitter sentiment and stock prices; Souza et al. (2016) similarly to Dickinson and Hu (2015) showed empirical evidence of Twitter sentiment predicting excess of log-returns for stocks of the selected retail companies. Google trends search data has shown to be useful for predicting unemployment claims, auto-mobile sales, travel destination planning and consumer confidence (Choi and Varian, 2012); for anticipating stock market moves (Preis and Moat, 2015, Moat et al., 2014, Preis et al., 2013); for improving nowcasting of suicide rates compared to results produced by models based on past suicide data alone (Kristoufek et al., 2016);

Previous research has also proven the role of web data for forecasting sales. For example, Gruhl et al. (2005) used blog data to predict book sales; Asur and Huberman (2010) predicted opening sales for movies based on Twitter chatter, while Liu (2006) were able to predict both aggregate and weekly box office revenue based on chatter from Yahoo Movies Web site; Chevalier and Mayzlin (2006) demonstrated that there is a strong relationship between consumer reviews and relative sales of books on Amazon; Kulkarni et al. (2012) demonstrated that online search data can be used as movies sales forecasting measure.

As it has been demonstrated, a vast amount of literature exists that shows the value of utilising Twitter, Facebook or Google information for predicting the future. One limitation of the existing studies is that they mainly focus on predictions of global events, such as elections, protests, disease outbreaks, or, on forecasting sales for popular products that naturally cause a lot of discussion, such as movies and books. The question whether online chatter can be relevant for predicting demand for general products remains open. Only few studies performed such analysis. For example, Dijkman et al. (2015) revealed that the relationship between tweets and sales did not hold for the products that did not receive a lot of social attention. The authors, nevertheless, were able to find some significant correlations after performing additional filtering of tweets based on their polarity, type of user and type of tweets. The impact of Facebook likes on sales was investigated by Lee et al. (2016): they showed that Facebook likes drive traffic and increase sales; Boldt et al. (2016) looked at quarter sales and Facebook data for Nike and found that Facebook data contains information about future sales; Du et al. (2015)

provided evidence that Facebook is positively associated with companies annual sales, however the impact of Facebook on sales is quite small; Liu et al. (2010) analysed online reviews and showed that their sentiment is correlated with the sales performance of products. While mentioned research works provided some evidence of the relationship between user-generated data and sales for general products, these studies have one significant limitation - they use estimated sales data (through Amazon or Groupon) and primarily analyse quarter or annual sales. Since the life cycle of retail products is very short and demand is changing within hours or days, the role of social media promises to be the most beneficial for short-term predictions, such as next-day or next-week sales. The literature on daily sales forecasts is almost absent, leaving a large space for further exploration.

Another limitation of the previous works is that the analysis is generally performed for only one type of online data at a time. To my best knowledge, only Bughin (2015) combined multiple sources of online data for predicting sales. Bughin (2015) used Twitter, Facebook, Google trends and blogs data, and explored the relationships between these social media variables and sales by employing the Vector Error Correction Model (VECM). The results of the study demonstrated that including social media variables into the model allowed to improve monthly telecom sales forecasts. However, the study was limited to the analysis of telecom industry in one country, thus, it cannot be generalised for other industries and geographies. The study also does not explore the impact of Google searches on daily sales. The latest study by Cui et al. (2017) assessed information value of Facebook, as well as advertising and promotion data for making predictions. Cui et al. (2017) showed that using social media information yielded improvements for daily sales forecasts in the out-of-sample accuracy ranging from 13% to 23%. They also provided evidence that non-linear models outperform linear models. However, the study was also limited to the analysis of one apparel retailer and only covered the period of 6 months. Other social media indicators such as Twitter and Google Searches trends were not included in the models.

In terms of forecasting techniques, the main methods that evolved in the last decades for modelling seasonal variations include exponential smoothing, time series regressions, autoregressive integrated moving average (ARIMA), artificial neural networks (ANN), and extreme learning machine (ELM) among others.

Multiple Linear Regression (MLR) is one of the most commonly applied forecasting techniques in the retail domain as it allows to consider simultaneously the impacts of multiple factors (Rogers, 1992). A much cited paper studying the propensity for social media metrics to predict sales is that of Asur and Huberman (2010). The authors used linear regression in order to forecast movie box-office revenues. Twitter metrics included volume of tweets about a movie, proportion of subjective tweets and whether the majority of tweets had positive or negative sentiment. Lassen et al. (2014) used multiple linear regression with twitter data to predict quarterly iPhone sales. Alon et al. (2001) demonstrated that Winters exponential smoothing and ARIMA models were viable methods in stable macroeconomic conditions. Ramos et al. (2015) showed that state space models can be very competitive compared to ARIMA models in producing automatic forecasts of univariate time series which are often needed in any retail business.

One limitation of the models mentioned above is that they are essentially linear models. However, the genuine relationship between the variables is often quite complex and rarely known, as shown by Cui et al. (2017), Souza and Aste (2016), for example. Thus, it is important to also explore non-linear models. One nonlinear method that has recently received extensive attention in forecasting is the artificial neural network (ANN) (Chu and Zhang, 2003). It has been shown that on average non-linear models in the form of neural networks are able to outperform traditional linear models in out-of-sample forecasting (Alon et al., 2001, Chu and Zhang, 2003). Franses and Draisma (1997) proposed to use ANNs when seasonal patterns change over time. Au et al. (2008) used the optimized ANN for fashion retail forecasting and showed that it outperforms the traditional SARIMA model. The interested reader can refer to the good review of the application of ANN for retail sales forecasting in (Zhang, 2009).

Another major limitation of many existing studies in forecasting is that they don’t perform stationarity checks for the variables. However, if sales and exogenous variables are non-stationary, which is often the case, the relationship between variables might be spurious (as pointed out by Granger and Newbold (1974), Yule (1926), Ventosa-Santaularia (2009)). To avoid spurious regressions, one should difference the data to make it stationary (as described in section 2.2.4). One important exception to this rule is when there is a cointegration between the non-stationary variables (section 2.2.5). In cases of cointegration a Vector Error Correction Model (VECM) can be used on levels of data (Cui et al., 2017). Bughin (2015) provided the evidence of cointegration between online information and telecom brands sales and showed that sales predictions based on the Error Correction Model (ECM) (Cui et al., 2017) were 25% more accurate than a simple auto-regressive model of sales; Iqbal and Uddin (2013) compared the forecasting performance of ECM with ARIMA and the VAR techniques and showed that ECM with macroeconomic data performed better for forecasting the monetary aggregate M2 for long run horizon; Jansen and Wang (2006) found that the forecasting performance of the ECM, arising from the cointegration relationship between the equity yield on the SP500 and the bond yield, was better relative to the univariate models for ten years forecast horizon.

In the current research work I aim to expand existing line of research by aggregating multiple sources of online data and comparing their ability to improve sales forecasts across multiple industries and multiple brands. I perform out-of-sample forecasting of aggregate retail sales using traditional linear models, the VECM to account for the long-run dynamic between cointegrated variables and the ANN to explore the nonlinear relationships between the time series.

2.4.2 Forecasting Techniques

In this section I provide basic information about the main types of forecasting techniques.

2.4.2.1 Qualitative Methods

Qualitative methods include forecasting approaches based on subjective estimates from individuals with vast domain knowledge, for example, visionary forecasting, market research or Delphi method. These methods are useful when no quantitative data is available. Whilst the extensive knowledge of experts might provide a good insight into future scenarios, bias in this approach, related to human judgement is very common (Lovallo and Kahneman, 2003). Qualitative methods, since they result from lengthy research and substantial time commitment from individuals with vast domain knowledge, also tend to be extremely expensive. This study will not concern itself with such techniques, as such domain expertise is not available, and furthermore, such forecasts are not replicable in future contexts.

2.4.2.2 Time Series Analysis and Projection Methods

Projection methods are based on using historical time series data, identifying patterns in it and then extrapolating those patterns forward into the future. This is an intuitive approach to forecasting which relies on the idea that behaviours in data have inertia, and are likely to continue repeating themselves. The examples of projection methods include Naive Method, Exponential Smoothing, Moving Average.

2.4.2.3 Casual Models

Whilst projection models and time series analysis fundamentally only look for repeating patterns in the time series, causal models try to identify factors which cause the changes in order to make predictions. These models are particularly useful when historical data is present for both, depended and independent variables. The most popular casual models include Regression Models, Econometric Models, Leading Indicators and Correlation Methods.

In this thesis I will focus on using casual models since the primary objective of the research is to analyse the ability of independent variables, such as social media indicators, to generate accurate forecasts. Following the literature review presented in the previous section, I will use Multiple Linear Regression as the benchmark, Artificial Neural Network as the model to account for potential non-linear relationships between variables and Vector Error Correction Model to account for cointegration. The description of the models is presented below.

Multiple Linear Regression. The general purpose of multiple linear regression is to examine how multiple independent variables are related to a dependent variable. For example, in this thesis the goal is to know how social media information about the brand (Twitter volume, Twitter sentiment, Facebook likes and posts, Google searchers) relates to sales of that brand. Mathematically, the dependent variable can be expresses as a linear combination of the explanatory variables as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon, \tag{2.36}$$

where y is the dependent variable, $x_1, \ldots, x_k$ are the predictive variables, β_0 is a regression constant or intercept, $\beta_1, \ldots, \beta_k$ are regression coefficients (each is the partial derivative of y with respect to that x), ϵ is an error term under assumption that it has normal distribution with mean 0 and constant variance.

The estimated OLS used for prediction of y has the form:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k, \tag{2.37}$$

where $\hat{y}$ is the expected value of y for a given set of x values, $b_1, \ldots, b_k$ are slopes estimates obtained by minimising the sum of squared errors:

$$\text{minSum} = \sum_{i=1}^{k} (y - \beta_0 - \beta_1 x_1 - \beta_2 x_2 - \ldots - \beta_k x_k)^2, \tag{2.38}$$

How well the equation fits the data is expressed by R^2, the "coefficient of multiple determination." It ranges from 0 (means there is no relationship between y and the x variables) to 1 (means there is no difference between the observed and expected values). For further details on MLR see Wooldridge (2008). In order to determine whether a particular x-variable is making a useful contribution to the model, one needs to examine the p-values for all coefficients in the model.

Vector Error Correction Model. One way to deal with non-stationary but cointegrated time series is to run MLR on differenced data. However, if one differences I(1) data, the long-run information will be lost. In order to capture both, short-run and long-run dynamics, one can use an Error Correction Model (Engle and Granger, 1987) for two variables or Vector Error Correction Model (Kilian and Ltkepohl, 2017) for more variables. The VECM links the short-run dynamic between time series with the long-run equilibrium relationship implied by cointegration. The mathematical representation is constructed based on the Johansen procedure by implicitly substituting Π in equation 2.25 for the lagged error correction term $\alpha\beta' y_{t-1}$:

$$\Delta y_t = \alpha\beta' y_{t-1} + \sum_{i=1}^{p-1} \Gamma_i \Delta y_{t-i} + u_t, \tag{2.39}$$

For further information on VECM refer to Kilian and Ltkepohl (2017).

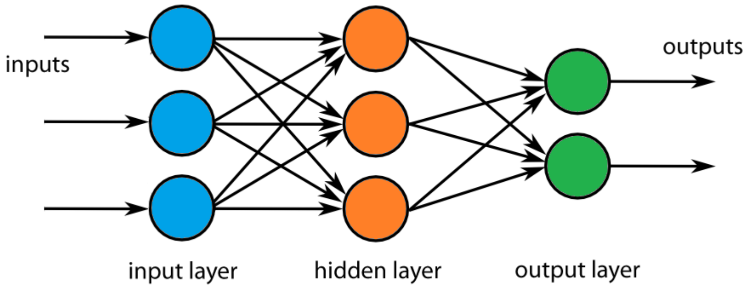


Figure 2.8: Graph representation of a feed-forward Artificial Neural Network Model.

Artificial Neural Networks. One of the major limitations of the previous two methods is that they are essentially liner methods. ANN is an advanced model which is able to learn complex non-linear relationships between input features and output variables. It processes information through differing weighted edges between nodes (artificial neurons), similar to the synaptic connections between biological neurons in the brain. As described in (Krenker et al., 2011) the information comes into the body of a neuron through weighted inputs. The body of an artificial neuron then sums the weighted inputs, bias and processes the sum with a transfer function before being sent out from the node. Mathematically, the process can be described as follows:

$$y(k) = F\left(\sum_{i=0}^{m} w_i(k)x_i(k) + b\right), \tag{2.40}$$

where $x_i(k)$ is input value in discrete time k, $W_i(k)$ is the weight, b is bias, F is a transfer function and $y_i(k)$ is output value in discrete time k.

Fig. 2.8 shows a representation of a simple multi-layer feed-forward ANN. The input vector consists of three nodes that are mapped to a hidden layer, which then gives two output values.

As discussed by Zhang (2009), before an ANN can be used for forecasting, it is important to determine the order of the network and the weights (parameters) of the model. The estimation of parameters is typically done on the training set using some candidate models. The model that performs the best on the validation set is selected. The out-of-sample observations are used to further test the performance of the selected model to simulate the real forecasting situations. The standard three-layer feed forward ANN can be used for time series forecasting in general and retail sales in particular.

ANN is a powerful model which is applicable to a wide range of problems, including time series prediction. Chu and Zhang (2003) investigated the power of ANN in predicting sales and demonstrated that ANN do better than linear models.

Logistic Regression Logistic Regression (LR) is well-suited for binary classification problems. In this study it is used to predict the direction of next day sales: ″up″ or ″down″. At the core of LR is the logistic function, also called the sigmoid function:

$$\sigma(t) = \frac{1}{1+e^{-t}} \tag{2.41}$$

If t is a simple linear function of explanatory variables $x_1,\dots,x_k$ then the logistic function can be written as:

$$y = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\dots+\beta_k x_k)}}, \tag{2.42}$$

where y is the predicted output, β_0 is the intercept and $\beta_1,\dots,\beta_k$ are the coefficients that can be estimated from the training data using maximum-likelihood algorithm. For the classification task the probabilities produced by logistic function can be mapped to a binary class value, using the following rule: 0 if $y < 0.5$; 1 if $y \geq 0.5$.

2.4.3 Evaluation of Forecasts

2.4.3.1 Cross Validation

In the case when a limited number of observations is available for validating a model, one can use the time series cross validation technique (as described by Hyndman (2016)). Cross validation approach uses many different training sets, each one containing one more observation then the previous one. For one-step ahead forecasts the schematic representation of the process is presented in Fig. 2.9.

If k observations are used to produce a forecast, then the 1-step-ahead forecasting works as follows (Hyndman, 2016):

1. Select the observation at time $k+i$ for the test set, and use the observations at times $1,2,\dots,k-1+i$ to estimate the forecasting model. Compute the 1-step error on the forecast for time $k+i$.

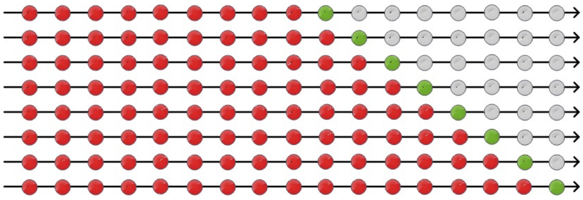


Figure 2.9: Schematic representation of iterations for one-step ahead forecasts. Each training set (red dots) contains one more observation than the previous one, and each test set (green dots) contains the next observation. The light gray dots are not used in the forecast.

2. Repeat the above step for $i = 1,2,\dots,T-k$ where T is the total number of observations.

3. Compute the forecast accuracy measures based on the errors obtained.

In the current thesis k was selected to be equal to 548, thus, the first training step was performed on the 70% of data.

2.4.3.2 Evaluation Metrics

In this thesis the analysis is performed for multiple brands that might have sales of varying magnitudes. For this reason, in this thesis I focus on the forecasting evaluation metrics that are not prone to changes in the magnitude of time series, thus, allowing to compare models performance across different brands. One of such metrics is Mean Absolute Percentage Error (MAPE). The functional representation of MAPE is:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}|\frac{y_i - \hat{y}_i}{y_i}| * 100 \tag{2.43}$$

where y_i is the observed value, $\hat{y}_i$ is the predicted value and n is the number of points to forecast. When serious outliers are present among the forecast errors, it is also recommended to report the Median Absolute Percentage Error (MdAPE) in addition to the MAPE, since medians are less sensitive than mean values to distortion from outliers.

In order to directly measure the improvement achieved by the proposed forecast method over the benchmark the Relative Mean Absolute Error (RelMAE) can be used which is calculated as follows (Shcherbakov et al., 2013):

$$\text{RelMAE} = \frac{\text{MAE}}{\text{MAE}^{\star}}, \tag{2.44}$$

where MAE and MAE* are the mean absolute error for the proposed forecasting model and the benchmark model, respectively. The MAE is calculated as:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{2.45}$$

When RelMAE < 1, the proposed method is better than the benchmark method, and when RelMAE > 1 the proposed method is worse than the benchmark method.

A comprehensive survey of a variety of forecasting evaluation techniques can be found at (R.J. and Koehler, 2006).

To test whether the forecasts from different models are statistically different from the benchmark model, the Diebold-Mariano test (Diebold and Mariano, 1995) can be used. In case when the goal is to select the best performing model among multiple algorithms, it is important to perform a joint analysis of the results achieved by various models in order to control the Family-Wise Error Rate (FWER). FWER is defined for the multiple hypothesis test as the probability of making one or more false discoveries among all the hypothesis. In this thesis the multiple hypothesis test was performed using a Friedman test with post-hoc Holm correction of the p-value (Derrac et al., 2011, Demšar, 2006).

2.5 Chapter Conclusions

During the process of literature review, few gaps in the fields of sentiment analysis, time series analysis, event detection and demand forecasting were identified that can be summarised as follows:

- The use of language in social media is constantly changing. There are not many studies that investigate the role of emoticons, abbreviations and slang for expressing emotions when communicating online.
- When analysing the role of social media for sales predictions, existing studies generally discus events or products that receive a lot of social attention, such as elections, disease outbreaks, protests, or sales for movies and books. Local events, such as spikes in sales for a company or sales for general products are not being investigated.

- A few studies that discuss the impact of social media on sales for ”general” products are limited to one or two companies only and a few geographies. This makes it impossible to generalise the findings of these studies across industries.

- Demand in retail industry is very volatile and the life cycles of products are short (Sen, 2008), however, the majority of research works in sales forecasting focus on quarter or annual sales data. The short term value of social media data, which promises to have the greatest impact, remains unexplored.

- Many social media indicators have been analysed, such as Twitter, Facebook, blogs, reviews, however, the analysis is generally done for one social media indicator at a time. There is space to explore the role of different social media sources and measure the difference in their predictive power.

- Analysis of social media events is predominantly done based on spacial, temporal and textual information using techniques from natural language processing. Research works that perform events clustering based on the nature of information propagation (for example, endogenous vs. exogenous events) are limited. I believe, the clustering of events using time series techniques could allow to account for different growth and relaxation signatures of events and could reveal new information about their internal dynamics.

- Studies on forecasting are often limited by the set of techniques used in a particular field, for example:

  - Research works that use econometric models such as Error Correction Models often don’t explore non-linear relationships between the variables.

  - Statistical approaches to the analysis of social media predominantly focus on the explanatory power of independent variables and don’t investigate the forecasting abilities of these variables on the out-of-sample datasets.

  - Data scientists use ”black box” models and often don’t perform stationarity and cointegration checks of variables, which might result in identifying spurious relationships.

In this thesis I aim to address the limitations of existing studies described above by using daily sales data for 75 brands, provided by Certona, and by developing a framework that allows to perform events study, stationarity and cointegration checks, and explore linear and non-linear relationships with the purpose of finding the model that performs best on the out-of-sample dataset.

Chapter 3

Data Collection and Pre-Processing

This chapter provides information about the types of data that were used for analysis in this thesis (see Fig. 3.1). The process of data retrieval and data filtering for each data type is described in detail. The chapter also discusses the approach for time series creation and time series pre-processing. Additionally, the chapter describes datasets that were used for sentiment model training and validation.

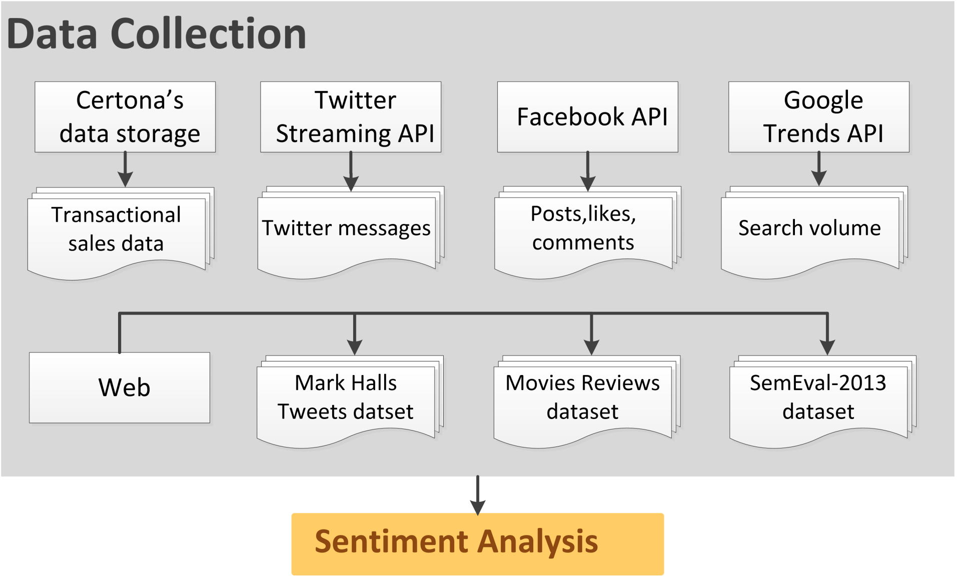


Figure 3.1: A framework for data collection.

3.1 Transactional Sales Data

In the research work I used time-stamped transactional sales data for 75 global brands, supplied by industrial partner Certona, and related to the sectors of apparel, footwear, body care, furnishings, games and services. For 51 of 75 brands the data covered the period of one year, from November 1, 2013, to October 31, 2014. For 24 of 75 brands the sales data covered the period of two years, from November 1, 2013, to October 31, 2015.

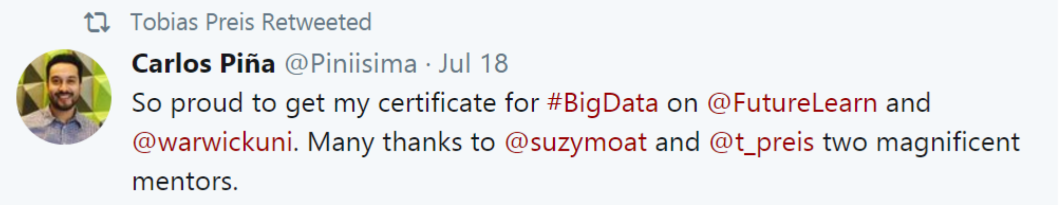
3.2 Twitter Data

3.2.1 Twitter Basics

Twitter [1] was launched on July 13, 2006, and during the following few years became one of the most popular online micro-blogging services and is now in the top 10 most visited internet sites. According to Statista[2] research as of August 2017, Twitter has over 387 million active users.

Twitter obtained such popularity due to its simplicity and functionality: a user can very easily and quickly share short messages with the rest of the world via mobile or PC. Many web pages and applications now have a Twitter button, which allows to share the content of website with friends simultaneously. The ease of sharing the information made Twitter an integral part of everyday life for many millions of people with over 500 million of tweets being posted every day, containing news, links to videos and articles, but also personal information about peoples lives, their likes and dislikes.

The information shared on Twitter is public, however, each individual user only sees the messages posted by his/her friends, or more precisely, people he/she follows. The messages called tweets, are limited to the length of 140 characters and have a specific structure. The best way to describe the structure of the tweet is to look at the example:



Tobias Preis Retweeted

Carlos Piña @Piniisima · Jul 18

So proud to get my certificate for #BigData on @FutureLearn and @warwickuni. Many thanks to @suzymoat and @t_preis two magnificent mentors.

Figure 3.2: Sample Tweet

The example on Fig. 3.2 is a good representative of a tweet, because it contains the most important parts of the tweet:

[1] https://twitter.com

[2] https://www.statista.com/

- Replies to other users or mentioning of other users. This can be achieved by adding the @ symbol before the user name. In the example above the author mentions @FutureLearn, @warwickuni, @suzymoat, @t_preis. The user names are strings up to 15 characters in length constructed of letters, numbers and underscore sign.

- A popular way of sharing useful information on Twitter is to post the message published by somebody else. This process is called retweeting. The example message above is a retweet by Tobias Preis of the message by Carlos Pina.

- In order to indicate the keywords in the message users have a tool called hashtag #. If a user wants his message to appear in the search results for a particular topic, he should precede the topic words in a message with a hashtag. For example, the sample message above with the hashtag near BigData will appear as a result of search for a word BigData.

  The application of hashtags became so popular, that people started being creative in using them. For example, hashtags are being used as a replacement of emoticons to express feelings and emotions: "I just broke your favourite cup! # kidding".

- Another very common part of a tweet is a link. This way short by its nature tweets, not only can communicate the message, but provide access to a much wider range of information through shared links.

Due to its reach to the public Twitter became a new space for promoting products, advertising deals and promotions. More and more businesses are creating official Twitter accounts to do marketing and spread information about sales to their customers.

3.2.2 Twitter API

One way of obtaining Twitter data is to collect it using Twitter Application Programming Interface (API). Even though there are limits on the amount of data, which can be collected through API, and the process requires time to accumulate the data, this approach is the most suitable for the research purposes and has been used for data acquisition in this research work. Twitter provides three types of API, which serve different purposes :

1. **REST API**. The REST API is a web service implemented using the principles of REST (Representational State Transfer) that allows developers to interact with Twitter and use most of its features. The is a limit of a maximum 180 requests to REST API per 15 mins. With the fact that 100 tweets can be retrieved during each request, the number of tweets which can be obtained through the **REST API** can achieve 1728 thousands a day. This number should be satisfactory for most of the research purposes, however REST API is not designed to allow extracting tweets based on the specific keywords, which often is the requirement.

2. **SEARCH API**. This type of the connection to Twitter allows querying the data based on the specific keywords. The Search API will return the tweets which were created during the last 6-9 days. The search is based on relevance rather than completeness, which means that one cannot rely that absolutely all tweets on the given subject were returned. If the priority is to search for completeness, the streaming API should be considered.

3. **STREAMING API**. Streaming API does not provide access to the past data, however, it allows accumulating data by sampling the tweets from a stream. The developers with data intensive needs can establish a persistent HTTP connection and receive tweets as they are being posted.

   Keywords based search is permitted with the restriction of 400 filter words and/or 5000 user identifiers. The streaming API's focus is in completeness of data, therefore one can expect to receive all existing tweets for the selected topic while the volume of tweets is within the specified limit.

   The following types of data streams are accessible from the Streaming API:

   (a) *"Sampled streams"*: delivers a random sampling of Tweets at a statistically valid percentage of the full 100% Firehose (full volume of tweets). The free access level to the sampled stream ("Spritzer") is set to approximately 1% of the full 100% Firehose.

   (b) *"Filtered streams"*: filtered streams deliver all the Tweets that match a selected filter (e.g. keywords, brands, products or geographical boundaries). Filtering limits the volume of tweets to 1% of the volume of the Firehose. For example, if one will be tracking a term that is not making up more than 1% of tweets in the public Firehose, he will get 100% of the tweets that match the search terms.

(c) *"Gardenhouse"*: is an increased access level, provides up to 10% of the volume of the Firehose.

All things considered, Streaming API "Filtered Streams" type of access was selected for the purpose of this thesis, because it allows obtaining data for an unlimited period of time and filter the data based on brands' names.

As it will be described later in more detail, the list of terms which is being used for streaming Twitter data in this dissertation has been comprised of almost 400 terms related to brands' names and their variations. It has been measured, that the average number of tweets received per day for all 400 filter words has not been exceeding 500,000 tweets. Taking into account the statistics from Statista research that around 500 million tweets are being posted every day, it is likely that the amount of data for 400 brands does not exceed the limit of 1% and that all tweets related to selected brands could be obtained.

3.2.3 Keywords Selection

In order to limit the streaming of data to specific brands of interest, I filtered the list of clients provided by an industrial partner Certona using the following steps:

- Some brands' names are represented by commonly used words, for example Apple, Avenue, Bernards, Aero. Incorporation of these brands' names into Twitter filtering list would return a lot of non-relevant tweets. In the current research work I excluded the brands with such names from the analysis;

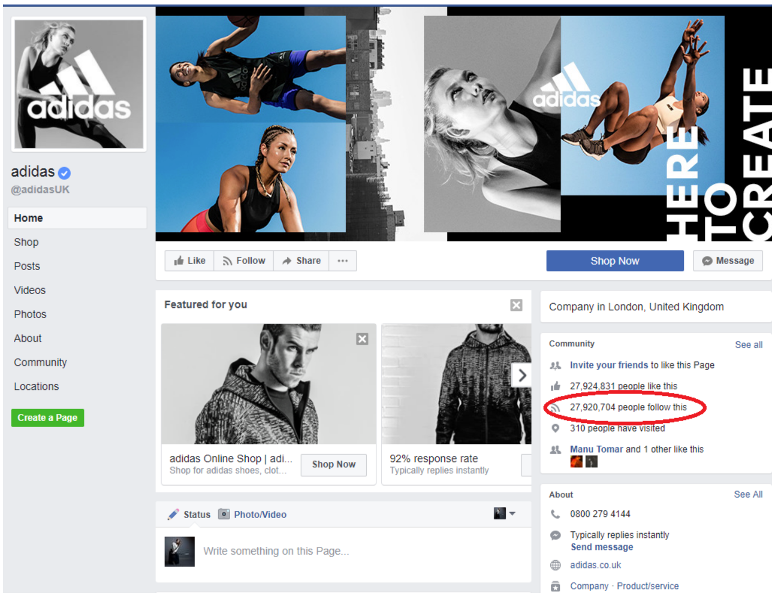
- In case when the brand name consisted of two or more words, for example Calvin Klein, the two variants, with the space between words (Calvin Klein) and without (CalvinKlein), were included in the terms list.

- Each brand name in the list was searched in Google to identify all the possible variants of the brand's name spelling. For example, for Victoria's Secret, 3 variants of how the brand's name could be written by customers, were identified: Victoria's secret, Victorias Secret, Victoriassecret. For the brand Toys "R" Us the identified variants were: Toys "R" Us, Toys"R"Us, "Toys"R"Us", ToysRUs. All the different variants of spellings were included in the terms list.

The final list of terms was comprised of 389 keywords. The statistical analysis of the tweets showed, that the terms definition approach described above allowed to gather more tweets about brands. For example, the term "Victoria's Secret" appeared 89,639 times in the tweets on the 20th of May 2013, however the spelling "Victoriasecret" also appeared that day, and allowed to collect additional 19,280 tweets about the same brand. "Calvin Klein" written with the space between two words was mentioned 32,117 times, while 4,240 people used spelling "CalvinKlein" when talking about the same brand.

3.3 Facebook Data

3.3.1 Facebook Basics



adidas Online Shop | adi...
Shop for adidas shoes, clot...

92% response rate
Typically replies instantly

Write something on this Page...

Company in London, United Kingdom

Invite your friends to like this Page
27,924,831 people like this
27,920,704 people follow this
310 people have visited
Manu Tomar and 1 other like this

0800 279 4144
Typically replies instantly
Send message
adidas.co.uk
Company · Product/service

Figure 3.3: A screenshot of Adidas UK official Facebook page. Circled in red is the number of page followers (almost 28 million).

Facebook[3] began in February of 2004 as a school-based social network at Harvard University. Since then, it has become the most popular social network in the world, increasing the number of daily active users by 18% each year and reaching over 2 billion active users in 2017 (Welch, 2017).

[3] https://facebook.com

Differently from Twitter, Facebook allows users to create a personal page on the web and use it for writing posts, sharing pictures and links to news or other interesting content on the Web, playing games, chatting live, and recently, even streaming live video. The content of each page can be made publicly accessible, or it can be shared only among a selected group of people (friends, family, fans). Some interactive Facebook features include the possibilities to commenting on other people's posts and pictures, or responding with emoticons, for example, likes. Users can also create events and invite their friends to participate.

One of the important reasons for Facebook's success is its attractiveness not only to people, but also to businesses. It became very popular for brands to have their official Facebook page and use it for communicating with their customers, providing content, sharing information about promotions and simply, for analysing customers opinions and understanding their needs. Fig. 3.3 shows a screen-shot of the official Adidas UK Facebook page. With the read circle I highlighted the number of page followers, which reached almost 28 people.

3.3.2 Facebook API and Data Filtering

Similarly to Twitter, Facebook developed an API that provides a possibility of accessing its historical data, Graph API. I worked in collaboration with a PhD colleague Giuseppe Pappalarado who developed a java script that allowed to download historical Facebook data using the Graph API. For the purpose of this research work, Giuseppe and I aimed to extract all Facebook pages associated with selected brands along with the number of likes, comments and posts that were registered for each page per day. Each page retrieved through the API had a hierarchical structure, as showed in Diagram 3.1.

Search function of the Graph API also provided several information fields for each page:

- Category: indicates which category the page is related to (i.e. Clothing, Local Business, Community, etc.);
- is_unclaimed: indicates whether the page was auto generated by Facebook and does not belong to any user, or whether it was generated by a registered user;

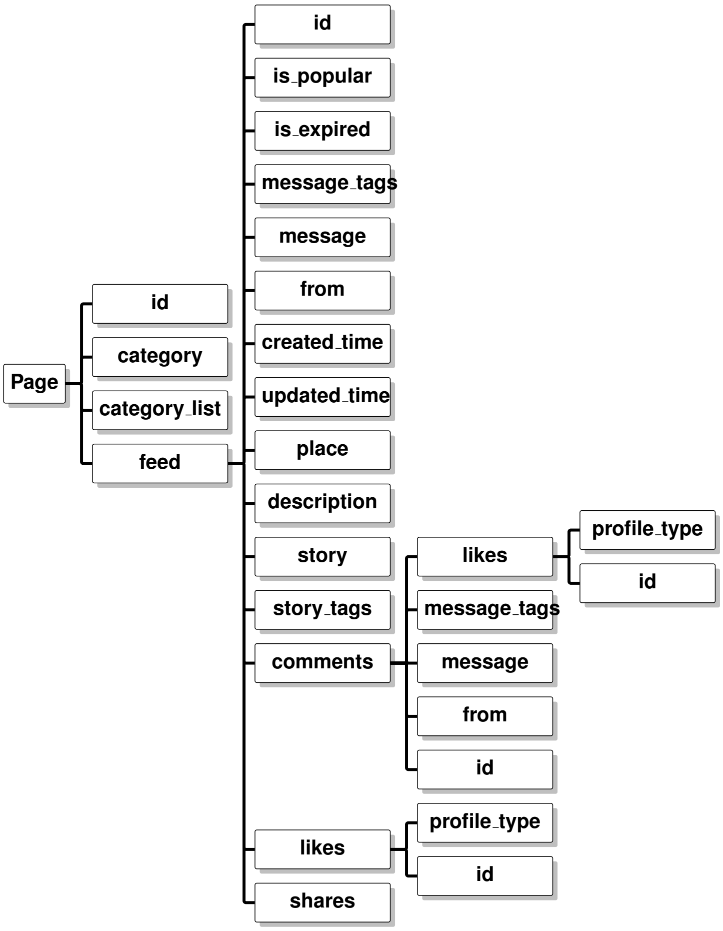


Diagram 3.1: Structure of the Data retrieved using Facebook Graph API for the entity Page.

- global_brand_root_id: a common id for multiple pages belonging to one brand;
- name: the main name showed on the page;
- website: the website associated with the page, if available;
- brand: the brand name used in the search;
- global_brand_parent_root: similar to *global_brand_root_id*, a name of the root page, common for multiple pages related to one brand;
- is_verified: indicates, whether the authenticity of the page owner has been verified by Facebook.

As the first step, Giuseppe Pappalarado found all Facebook pages that had "brand" field equivalent to brands' names of my interest. After examining queries results he discovered that each brand was associated with dozens or even hundreds of pages (see Table 3.1, column “Total Number of Pages”). Since a page on Facebook can be created by people, businesses or be automatically generated by Facebook for organisations, not all of the returned pages would be relevant for the task at hand. Furthermore, there are plenty of “fake” pages that people create in order to gain traffic/visitors, while pretending to be a popular brand. In order to identify the “genuine” pages related to brands, as the second step, Giuseppe and I selected only pages that satisfied the following filtering rules:

- Pages had to mention the official brands’ website in the “website” field;
- Pages had to to share *global_brand_root_id* or *global_brand_parent_root* with other selected pages;
- Pages had to be related to one of the preselected categories (Company, Shopping/Retail, Clothing, Health/Beauty, Product/Service, etc.). In order to create a categories list I extracted the unique categories from the non-filtered list of pages and manually excluded categories that had no relation to business, such as Lawer, Photographer, Author, etc.

The final number of selected pages for some of the brands are presented in Table 3.1, column “Filtered Pages”.

3.4 Google Trends Data

3.4.1 Google Trends Basics

Since the Google search engine is dominating the internet search market, statistical data about search keywords can be very insightful. Google Trends[4] is a tool that allows to obtain the popularity of search terms on Google (see Fig. 3.4 for visual interface of Google Trends website). Google Trends can be extremely useful for understanding the hottest search trends of the moment, along with those developing in popularity over time. It provides a number of features to narrow the search, such as possibility to specify geographic location, industries category, time periods of interest, and type of search (for example, image search, or web search). Google Trends also allows users to compare the volume of searches between two or

[4] `https://trends.google.co.uk/trends`

Table 3.1: Number of Facebook pages for some brands before and after filtering.

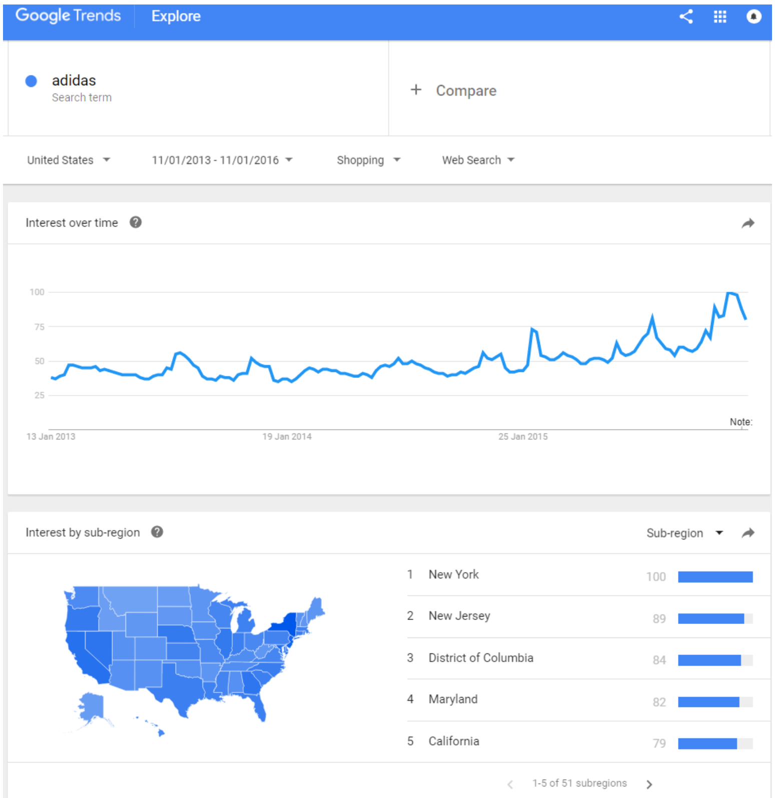
Brand	Total Number of Pages	Filtered Pages
Abercrombie	432	30
Adidas	493	89
Calin Klein	406	25
Chico's	447	202
Crocs	479	40
Gamestop	384	4
HomeDepot	421	10
kenneth cole	325	7
LaCoste	513	25
Levi	504	15
Loreal	504	101
Mattel	510	21
Meijer	417	50
newbalance	260	28
Nike	501	66
nissen	476	17
petco	366	17
Puma	510	30
radioshack	423	31
Reebok	427	114
Spanx	91	6
Speedo	471	36
Timberland	443	47
Tommy Hilfiger	411	37

more terms. Depending on the types of searches one performs, he/she can obtain a good proxy for the public's interests, concerns or intentions.

3.4.2 Google Trends Data Retrieval

Google Trends tool allows to download weekly data for a certain keyword from 2004 to 2015 as a single .csv file. However, daily data can only be downloaded in blocks of 90-day periods. This presents a problem, since the number of queries in each downloaded block is not a raw number of queries, but a rescaled value between 0 and 100, representing the relative popularity of a keyword over selected period of time. For consistency, the data from each 90-days period needs to be rescaled to one scale and then stitched together into one file.

To extract Google Trends searches data related to brands I used a Python script developed by Clinton (2014). The script allows to retrieve data programmatically by specifying search terms, granularity and time periods of interest, and it also



adidas
Search term

United States · 11/01/2013 - 11/01/2016 · Shopping · Web Search

1	New York	100
2	New Jersey	89
3	District of Columbia	84
4	Maryland	82
5	California	79

Figure 3.4: Example of Google Trends data visualisation for Adidas search term in the Shopping category.

automatically rescales and stitches daily data.

It is important to notice that Google search data results are based on a sample, rather than on the full population of search requests. This should not create a problem for common search terms, however, searches with a low overall volume can produce misleading results.

For the purpose of this research work I used the names of brands as the search terms and extracted daily searches time series for each brand, for the period from November 1, 2013, to October 31, 2015.

3.5 Time-Series Creation

During the first stage of this research work, sales and social media data was retrieved as discussed in section 3. During the second step, I performed sentiment analysis on Twitter data to extract additional information about the valence of messages (the process of sentiment extraction will be described in section 4). In order to perform any analysis on the collected data, such as modelling, forecasting, events study, it was necessary to accumulate data into time series at a specific frequency. In this study I used 24-hour aggregation window since I was interested in identifying large enough social media events that last for at least one day or more. This duration of social media spikes would have practical implications for business managers, giving them enough time to react to changing consumers demand. In this study I did not aim to explore larger aggregation windows such as 1 week, since the response on social network is very dynamic, and it is very likely that the impact of social media spike would dissipate after 1-2 weeks. It would also be more difficult to drive conclusions about the true relationship between spikes in sales and spikes in social media as the time goes by.

Using the data obtained through data collection process and sentiment analysis, the following time series were created to be used in further analysis:

1. **Sales volume time series**. The sales transaction dataset contained accumulation of time-stamped transactional records across multiple products related to each brand that has been recorded at no particular frequency. To create sales time series for each brand I aggregated the revenue across different products within the 24-hour time interval individually for each brand.

2. **Tweets volume time series**. To create tweets volume time series for each brand I calculated the number of tweets that were posted posted on every day for each brand.

3. **Volume of positive tweets time series**. Similarly to the volume of tweets, I calculated the number of positive tweets that were posted posted on every day for each brand.

4. **Volume of negative tweets time series**. Represents the number of negative tweets that were posted posted on every day for each brand.

5. **Tweets sentiment time series**. This time series represents daily sentiment score calculated as the ratio of number of positive tweets to a sum of positive

and negative tweets:

$$\text{sentScore} = \frac{\text{count}(\text{PosTweets}_{\text{brand}})}{\text{count}(\text{PosTweets}_{\text{brand}}) + \text{count}(\text{NegsTweets}_{\text{brand}})} \tag{3.1}$$

6. **Facebook volume of posts** Since sales data was aggregated across different products per brand, I also aggregated Facebook data across different pages. Facebook volume of posts represents the daily number of posts across all pages relevant to a specific brand (allows to measure brands' engagement).

7. **Facebook volume of likes**. Represents the daily number of likes across all pages (allows to measure brands' sentiment).

8. **Facebook volume of comments**. Represents the daily number of comments across all pages (allows to measure brands' awareness).

9. **Google Trends search volumes**. Represents the daily number of searches for a brand scaled between 0 and 100.

3.6 Data Pre-Processing

Before time series could be used for analysis it was required to perform some pre-processing steps as described in the sections below.

3.6.1 Missing Values Imputation

I conducted exploratory analysis of datasets and excluded brands that had more than 10% of data missing for any of the dependent and independent variables. For the remaining brands, I performed data imputation for the periods with missing data. To preserve the weekly variation, I imputed each missing data point with the weighted average of the previous data points that belonged to the same day of the week as the missing point. To give more weight to the points that appeared closer in time we used inverse distance weighting (Shepard, 1968):

$$x_t = \frac{\sum\limits_{i=1}^{N} \frac{1}{d(x_t, x_{t-7i})^p} x_{t-7i}}{\sum\limits_{i=1}^{N} \frac{1}{d(x_t, x_{t-7i})^p}}, \tag{3.2}$$

where x_t is the value to be imputed, N is the number of points to be taken into account for imputation (5 in this study), d is the distance operator (in this study it is

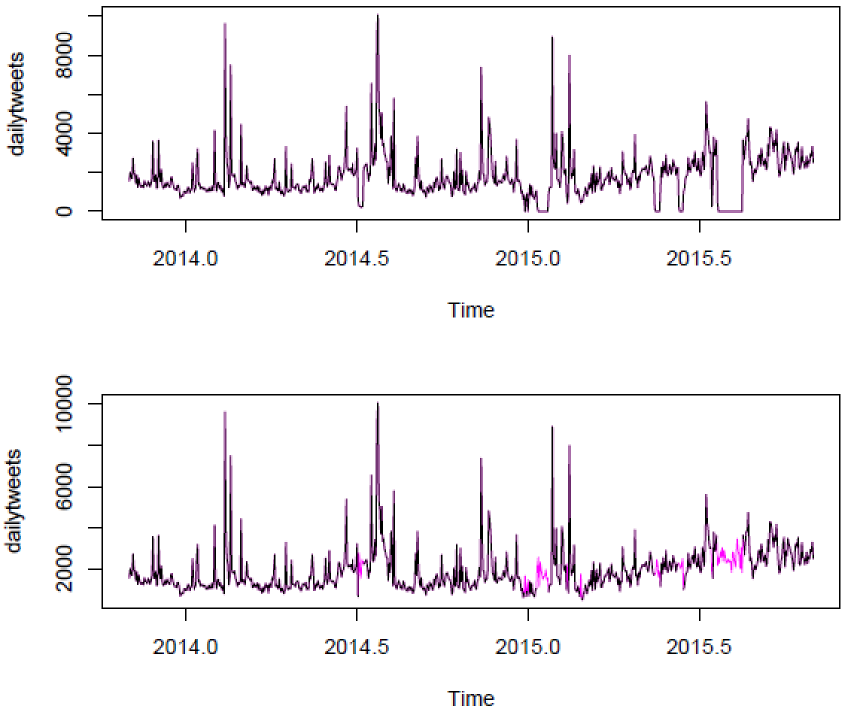


Figure 3.5: a) An example of a brand with periods of missing Twitter volume data; b) An example of missing Twitter volume data being imputed as a weighted average of previous values.

a number of days between the known point x_{t-7i} and the unknown point x_t), and p is the power parameter (2 in this study) that controls the influence of the distance (the higher is p, the greater is the influence of values closest to the interpolated point).

An example of a brand with missing Twitter volume data and imputed data can be seen in Fig. 3.5.

3.6.2 Data Transformation

I also performed the log transformation of sales and social media data, as it allowed to induce comparability in data with values that range over several orders of magnitude:

$$x' = \log(x) \tag{3.3}$$

Further, the data was normalized by using z-score Norman and Streiner (2008):

$$x' = \frac{|x - \overline{X}|}{\sigma}, \tag{3.4}$$

where $\overline{X}$ is mean and σ is standard deviation. As a result, all variables in the dataset had equal means of 0 and standard deviations of 1 (unit variance). It is important to notice that mean and standard deviation were calculated only from the values of the training set. This is to make sure that out-of-sample set used for training had no "future information" incorporated in it by standardization procedure. After the forecasts were produced the data was denormalised and delogged in order to obtain sales value of true magnitude.

3.7 Other Relevant Datasets

To perform sentiment classification labelled training and test sets are required. The quality and the size of the dataset have a strong impact on the final quality of classification. Manual classification of thousands of tweets would be a long a tedious process. For the purpose of this research I used publicly available datasets described below.

3.7.1 Movies reviews sentences Dataset

Movies reviews sentences dataset is a corpus based on the dataset introduced by Pang and Lee (2005) and consists of individual sentences extracted from movies reviews. The dataset contains 5,331 positive and 5,331 negative snippets. Some example sentences from the dataset are presented in Table 3.2.

Table 3.2: Examples of positive and negative reviews from the movies reviews dataset.

Positive Reviews	Negative Reviews
"a feel-good picture in the best sense of the term"	"stinks from start to finish, like a wet burlap sack of gloom"
"engagingly captures the maddening and magnetic ebb and flow of friendship"	"offers absolutely nothing I hadn't already seen"

3.7.2 Mark Hall's Dataset

Mark Hall is a core developer of Weka data mining software. He has released a labelled datasets of tweets Hall (2012). Tweets were classified into two classes using the variety of different techniques. The dataset is comprised of 42,581 positive tweets and 8,599 negative tweets. The the dataset has the following format:

[tweetId], [text], [class(pos/neg)]

Examples of tweets:

1229417433, ′youŕe so sweet! proving me right again... the Dutch are the Best! ′, pos

1230820258, ′my neighbors r complaining about my home theater ...so nights i have to switch off the subwoofer ′, neg

3.7.3 Semeval-2013 data

SemEval (Semantic Evaluation) is an international workshop that focuses on the evaluation of semantic analysis systems. It was created with the purpose of encouraging research that helps to understand the dynamic of conveying sentiment in short messages, in particular tweets and SMSs. My interest lied in Task 2-B: Sentiment Analysis in Twitter, a message-level classification:"Given a message, decide whether it is of positive, negative, or neutral sentiment. For messages conveying both a positive and a negative sentiment, whichever is the stronger one was to be chosen" (Nakov et al., 2013). The organisers of the workshop (Nakov et al., 2013) created the SemEval Tweet corpus which is comprised of Twitter messages annotated with the message-level polarity. The dataset was broken into Training, Development and Test groups and released to the research community. The training dataset is comprised of 3,662 positive tweets, 1,466 negative tweets and 6,600 neutral tweets. The test dataset is comprised of 1,573 positive tweets, 601 negative tweets and 1,640 neutral tweets.

Chapter 4

Twitter Sentiment Analysis

This chapter presents a step-by-step approach for two main methods of sentiment analysis: lexicon based approach and machine learning approach. I show that accuracy of the sentiment analysis for Twitter can be improved by combining the two approaches: during the first stage a lexicon score is calculated based on the polarity of the words which compose the text, during the second stage a machine learning model is learnt that uses the lexicon score as one of the features. The results showed that the combined approach outperforms the individual approaches. I demonstrate the use of the algorithm on a combined tweets and movies reviews dataset and on popular dataset from a Twitter sentiment competition SemEval-2013, task 2-B. Background and literature review for this chapter are presented in section 2.1

4.1 Introduction

Sentiment analysis is an important part of Twitter analysis as it has been shown that sentiment might be a strong predictor of future consumer demand (Asur and Huberman, 2010, Liu et al., 2010, Chevalier and Mayzlin, 2006). The goal is to understand whether tweets about selected brands are positive, negative or neutral and find the relationship between tweets sentiment and sales data, as well as to identify the events which triggered the change of customers' mood.

4.2 Methodology

The sentiment analysis performed in this study is comprised of two components: lexicon-based algorithm and machine learning algorithm. In this chapter the process of implementing both algorithms is presented using the methodology described in detail in sections 2.1.3 and 2.1.4, respectively. The lexicon-based approach is based on using the dictionaries of words with predefined polarities and comparing the words in the sentence under analysis against the dictionary. The main advantages of the lexicon approach are that it is fast, easily interpretable and does not require a training set with labelled sentences. Also, the outcome of classification depends only on the lexicon and does not depend on the words/patterns that have been seen in the training set. This opens a possibility of extending the lexicons or replacing them with new domain specific lexicons without the need to retrain the model. The machine learning method is based on building feature vectors from each sentence and using supervised machine learning algorithms to train classification models. Machine learning algorithms allow to include additional features that might have valuable information about sentiment, such as number of positive/negative words; presence/absence of emoticons, presence/absence and the number of different parts of speech, among many others. Machine learning classifiers can identify patterns in the training data and are often more accurate that the lexicon method, however, the accuracy depends on the quality of the training set.

In this chapter I propose an algorithm that allows to capture the advantages of both algorithms by creating a hybrid method. As the first step after pre-processing, I performed the lexicon-based classification and store the obtained sentiment as the lexicon score. In the second step, I trained a machine learning model and use the lexicon score from the previous step as one of the features in the feature space. A framework for the proposed algorithm is presented in Figure 4.1.

The main advantage of the proposed approach that uses a symbiosis of lexicon and machine learning algorithms, is the ability to attain the best of both worlds, interpretability from a carefully designed lexicon, and the improved accuracy achieved by identifying patterns in data using supervised learning algorithms. Also, the combination of methods allows to implicitly include sentiment information about words that are present in the lexicon, but have not been seen in the training set.

Training and evaluation of sentiment models was performed for 2-classes and 3-classes classification problem using the following datasets:

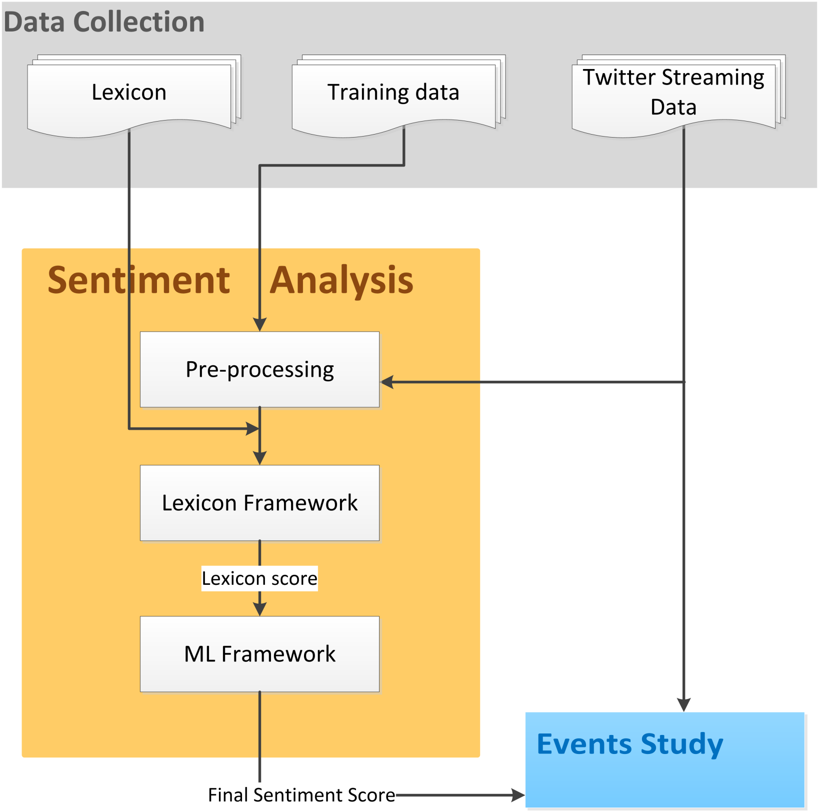


Figure 4.1: A framework for the hybrid sentiment analysis algorithm.

1. **A combined tweets and movies reviews dataset.** In this scenario I used Tweets dataset collected by Mark Hall (section 3.7.2). 80% of tweets from this dataset were used for generating a new lexicon in the lexicon approach or for training a classification model in the machine learning approach. The remaining 20% were used for testing. In order to have a larger test set I enhanced the remaining 20% of tweets from Mark's dataset with a Cornell movies reviews dataset (section 3.7.1). Since there were more positive than negative sentences in the dataset, I selected the same number of positive sentences as there was a number of negative sentences to create balanced training and test sets. The composition of the final dataset is presented in Table 4.1.

Table 4.1: A composition of a combined tweets and movies reviews dataset.

	Training set	Test set
negative	6879 tweets	1720 tweets + 5331 reviews
positive	6879 tweets	1720 tweets + 5331 reviews

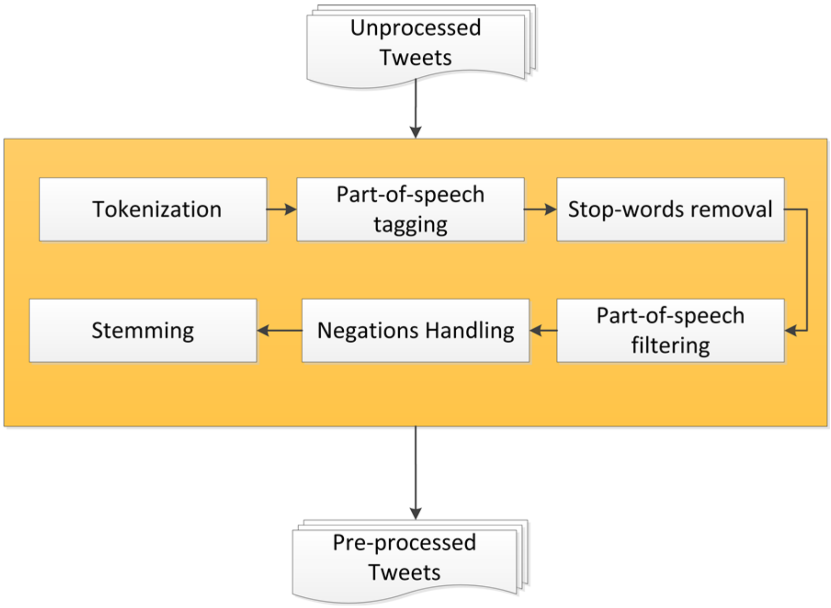


Figure 4.2: Pre-processing steps for Twitter data.

2. **A dataset from SemEval-2013 Competition, Task 2-B**. SemEval-2013 dataset Nakov et al. (2013) consists of separate training and test sets that are comprised of positive, neutral and negative messages. More information on the dataset can be found in section 3.7.3.

A detailed description of the sentiment analysis and evaluation process is presented in the following sections.

4.3 Twitter Data Pre-processing

Both, lexicon method and machine learning approach require data preparation stage or pre-processing. I perform the pre-processing before the actual methods of sentiment analysis are applied following the steps visualised in Figure 4.2. Detailed information on each step of pre-processing is provided in section 2.1.2.

Most of the pre-processing steps are performed using the WEKA[1] API for Java. WEKA was developed in the university Wakaito and provides implementations of many machine learning algorithms. Since it is an open source tool and has an API, WEKA algorithms can be easily embedded within other applications.

Tokenization. Tokenization is a process of creating a bag-of-words from text. For this purpose I used a tokenizer from a library called ArkTweetNLP[2], which was

[1] `http://www.cs.waikato.ac.nz/ml/weka/`
[2] `http://www.ark.cs.cmu.edu/TweetNLP`

developed by the team of researchers from Carnegie Mellon University and was specially designed for working with twitter messages. ArkTweetNLP recognizes specific to Twitter symbols, such as hashtags, at-mentions, retweets, emoticons, commonly used abbreviations, and treats them as separate tokens Gimpel et al. (2011). An example of the tokenization process is presented in Table 4.2, where the first column on the left is a column of tokens obtained by tokenizing the sentence.

Part-of-Speech Tagging (POS). Part-of-speech-tag is a process of tagging each token in terms of part of speech it belongs to: noun, pronoun, adverb, adjective, verb, interjection, intensifier among others. As described in section 2.1.2, POS might be an important feature in machine learning classification used for uncovering underlying patterns in language structure. In this study I explored multiple existing Pos-taggers: Stanford Tagger[3], Illinois Tagger[4], OpenNLP[5], LingPipe POS Tagger[6], Unsupos[7], ArkTweetNLP[8], Berkeley NLP Group Tagger[9]. It appears that all of them, apart from ArkTweetNLP, have trained their models on the news-type of documents. Since the performance of the POS tagger deceases on the out of domain data, I chosen to use a tagger from ArkTweetNLP Gimpel et al. (2011), because it is the only tagger from those mentioned above that was trained on a Twitter dataset. ArkTweetNLP authors developed 25 POS tags, with some of them specifically designed for Twitter symbols (see Table 4.3 for some tags examples). An example of how ArkTweetNLP tagger works in practice is presented in Table 4.2.

Stop-words Removal. During this step, words such as prepositions and articles were removed, since they don't carry sentiment. WEKA provides a file with a list of stop-words that can be adjusted to ones needs. In this study I used an original WEKA stop-list file without modifications.

Part-of-Speech filtering. Some of the parts-of-speech are irrelevant for sentiment classification. Examples include proper nouns (America, Mark, HSBC), URLs, email addresses, numerals and dates among others. Excluding such tokens

[3] http://nlp.stanford.edu/software/index.shtml

[4] http://cogcomp.cs.illinois.edu/page/software_view/3

[5] http://opennlp.sourceforge.net/models-1.5

[6] http://alias-i.com/lingpipe/demos/tutorial/posTags/read-me.html

[7] http://wortschatz.uni-leipzig.de/~{}cbiemann/software/unsupos.html

[8] http://www.ark.cs.cmu.edu/TweetNLP

[9] http://nlp.cs.berkeley.edu/Software.html

Table 4.2: Example of ArkTweetNLP tokenizer and a POS-tagger in practice (reprinted and adjusted by permission of the authors (Gimpel et al., 2011)). The first line is a sentence under analysis, the first column is a column of tokens extracted from a sentence by a tokenizer, the second column is a column of POS-tags assigned to tokens by a POS tagger, the third column is tag description and the last column is the confidence of tags prediction.

“ikr smh he asked fir yo last name so he can add u on fb lololol”			
Token	Tag	Description	Confidence
ikr	!	interjection	0.8143
smh	G	abbreviation	0.9406
he	O	pronoun	0.9963
asked	V	verb	0.9979
fir	P	preposition	0.5545
yo	D	determiner	0.6272
last	A	adjective	0.9871
name	N	noun	0.9998
so	P	preposition	0.9838
he	O	pronoun	0.9981
can	V	verb	0.9997
add	V	verb	0.9997
u	O	pronoun	0.9978
on	P	preposition	0.9426
fb	ˆ	proper noun	0.9453
lololol	!	interjection	0.9664

“ikr”	means “I know, right?”, tagged as an interjection.
“so”	is being used as a subordinating conjunction, which coarse tagset denotes P.
“fb”	means “Facebook”, a very common proper noun (ˆ).
“yo”	is being used as equivalent to “your”; our coarse tagset has posessive pronouns as D.
“fir”	is a misspelling or spelling variant of the preposition for.
	Perhaps the only debatable errors in this example are for ikr and smh (“shake my head”): should they be G for miscellaneous acronym, or ! for interjection?

from the sentiment analysis can provide additional type of filtering. In this study I only kept the following parts-of-speech: N(common nouns), V(verbs), A(adjectives), R(adverbs), !(interjectiosn), E(emoticosn), G(abbreviations, foreign

Table 4.3: Example POS tags developed in ArkTweetNLP (Gimpel et al., 2011).

Tag	Description
@	at-mention, identifes another user as recipient of the tweet
U	URL or email address
#	Hashtag, identifies the topic/category for a tweet
~	Discourse marker, indicates a continuation of a message across multiple tweets
E	Emoticons
G	Abbreviations, foreign words, garbage

Table 4.4: A list of negation words used during the negation handling step.

wouldn't
wouldnt
won't
wont
without
weren't
werent
wasn't
wasnt
shouldn't
shouldnt
shant
oughtn't
oughtnt
n't
nowhere
don't
dont
doesn't
doesnt
doesnot
didn't
didnt
daren't
darent
can't
cant
cannot
arn't
arnt
aren't
arenot
lacking
lack
isn't
isnt
isnot
havn't
havnt
haven't
havent
havenot
hasn't
hasnt
hasnot
hardly
hadn't
hadnt
nothing
not
nor
none
nobody
no
never
neither
needn't
neednt
mustn't
mustnt
mightn't
mightnt
lacks

words, possessive endings).

Negations Handling. During this step words that can change sentiment orientation of the following words were detected. The list of negation words used in the current study is presented in Table 4.4.

The process of negation handling is slightly different for the lexicon-based and machine learning approaches. In the lexicon-based pre-processing, the negation handling is implemented using simple, but effective strategy: if negation word is found, the sentiment score of every word, appearing between a negation and a clause-level punctuation mark (.,!?:;), is reversed (Pang et al., 2002). For example, before negation handling, the scores in the sentence below would be as follows:

*"These trousers are not comfortable **[+1]**, but I must admit the color is great [+1] and they are good [+0.45] value for money."*

After the negation step is performed the scores are modified:

*"These trousers are not comfortable **[-1]**, but I must admit the color is great [+1] and they are good [+0.45] value for money."*

There are, however, some grammatical constructions in which a negation term does not have a scope. Some of these situations were implemented in this study as exceptions:

Exception Situation 1: Whenever a negation term is a part of a phrase that does not carry negation sense, I considered that the scope for negation was absent and the polarity of words was not reversed. Examples of these special phrases include "not only", "not just", "no question", "not to mention" and "no wonder".

Exception Situation 2: A negation term does not have a scope when it occurs in a negative rhetorical question. A negative rhetorical question is identified by the following heuristic: (1) It is a question; and (2) it has a negation term within the first three words of the question. For example:
"Did not I enjoy it?"
"Wouldn't you like going to the cinema?"

In the machine learning approach word do not have predefined scores, therefore, in order to indicate the presence of negation, I modified the words, appearing between a negation and a clause-level punctuation mark, by adding a suffix _NEG to the end of those words. For example, the sentence from the previous example was be modified into:
"These trousers are not comfortable_NEG, but I must admit the color is great and they are good value for money."

This modification is important, since each word in a sentence serves a purpose of a feature in the machine learning approach. Words with _NEG suffixes increase the dimensionality of the feature space, but allow the classifier to distinguish between words used in the positive and in the negative context.

Stemming. The overall impact of stemming depends on the dataset and stemming algorithm. WEKA contains implementation of a SnowballStemmer Porter (2002) and LovinsStemmer Lovins (1968). After testing both implementations I discovered that the accuracy of the sentiment classification decreased after applying both stemming algorithms, therefore, in the final implementation of the sentiment analysis algorithm stemming operation was excluded.

4.4 Sentiment Analysis using Lexicon Approach

4.4.1 Lexicons Data

Lexicon-based approach for sentiment analysis requires dictionaries of words annotated with the word's semantic polarity. Dictionaries that were used for sentiment classification in this study are the following:

1. **Opinion Lexicon Hu and Liu (2004) (OL)**. This is a classic sentiment dataset comprised of 2,006 positive and 4,783 negative English words and compiled over many years. It is interesting to notice that some of the words in the dataset are misspelled, which did not happen by mistake. These words were included deliberately as these misspelled words appear frequently in the social media content. In order to use in my analysis I assigned a score of "+1" to positive words, and a score of "-1" to negative words. Example words from the opinion lexicon along with their polarities are presented in Table 4.5.

Table 4.5: Example words from the opinion lexicon Hu and Liu (2004) along with their polarity.

Positive Words	Score	Negative Words	Score
magic	1	afraid	-1
magical	1	aggravate	-1
magnanimous	1	aggravating	-1
magnanimously	1	aggression	-1
magnificence	1	aggressive	-1
magnificent	1	aggressiveness	-1
magnificently	1	aggressor	-1
majestic	1	aggrieve	-1
majesty	1	aggrieved	-1
manageable	1	aggrivation	-1
maneuverable	1	aggravation	-1
marvel	1	aghast	-1

2. **Collection of emoticons, social media slang and abbreviated words (EMO)**. Hogenboom et al. (2013) showed that incorporation of the emoticons into lexicon can greatly improve the accuracy of classification. Apart from emoticons, new slang words and abbreviations are constantly emerging and need to be accounted for when performing sentiment analysis. However, most of the existing public lexicons do not contain emoticons and social media slang, on the contrary, emoticons and abbreviations are often being

removed as typographical symbols during the first stages of pre-processing. As part of sentiment analysis, I aimed to explore the role of emoticons, as well as the impact of most popular abbreviations and social media slang words, on opinion expression. I examined online resources for commonly used emoticons, slang and abbreviations and constructed a lexicon of 242 positive and 162 negative tokens. Example tokens from the constructed EMO lexicon are presented in Table 4.6.

Table 4.6: Example tokens from EMO lexicon along with their polarity. Tokens represent emoticons, abbreviations and slang words that are commonly used in social media to express emotions.

Positive Token	Score	Negative Token	Score
l-)	1	[-(	-1
:-}	1	T_T	-1
x-d	1	:-((	-1
;;-)	1	:-[	-1
=]	1	:(((	-1
lol	1	dbeyr	-1
fab	1	iwiam	-1
ilum	1	nfs	-1
fbow	1	h8ttu	-1
iwalu	1	gtfo	-1
koc	1	wtf	-1

3. **A Lexicon Constructed from Trained Data (AUTO)**. I created a new lexicon of words with assigned to them polarity, specifically designed for sentiment analysis of Twitter messages. I used the approach of automatic lexicon generation from trained data described in section 2.1.3. The goal was to understand whether automatically generated lexicons could provide enough accuracy compared to the manually labelled lexicons. I used a dataset of Twitter messages produced by Hall (2012) that is comprised of labelled 41,403 positive Twitter messages and 8,552 negative Twitter messages. The method to generate a sentiment lexicon was implemented as follows:

   - Pre-processing: creating a bag-of-words by tokenizing the sentences in the dataset, assigning POS tags to all tokens in the bag-of-words, filtering irrelevant POS (stop-words, URLs, @-mentions, etc.).

   - Calculating the number of occurrences of each word in positive and negative sentences.

- Calculating the ”positiveness”, or positive score, of each word by dividing the number of occurrences in positive sentences by the number of all occurrences:

$$\text{positiveSentScore} = \frac{\text{Positive sentences}}{(\text{Positive sentences} + \text{Negative sentences})} \tag{4.1}$$

Similarly, the negative score was calculated by dividing the number of occurrences in negative sentences by the total number of mentions:

$$\text{negativeSentScore} = \frac{\text{Negative sentences}}{(\text{Positive sentences} + \text{Negative sentences})} \tag{4.2}$$

For example, the word “*Pleasant*” appeared 122 times in positive sentences and 44 times in negative sentences. According to Eq. 4.1 and Eq. 4.2, the positive score of the word “*Pleasant*” is:

$$\text{positiveSentScore} = \frac{122}{(122+44)} = 0.73,$$

and the negative score is:

$$\text{negativeSentScore} = \frac{44}{(122+44)} = 0.27$$

The positive score of the word “*Pleasant*” is higher than its negative score indicating its overall positive sentiment.

- Mapping the positive score into the range [-1; 1] by using the following formula:

$$\text{PolarityScore} = 2 * \text{positiveSentScore} - 1 \tag{4.3}$$

Mapping of the score in the range [-1; 1], where ”-1” indicates strong negative sentiment, ”0” indicates neutrality and ”+1” indicates strong positive sentiment, makes the words from AUTO lexicon comparable with the words from other lexicons.

Table 4.7: Example of sentiment scores of words in the automatically generated lexicon.

Word	Positive Score	Negative Score	Polarity Score
Good	0.675	0.325	0.35
Excellent	0.925	0.075	0.85
Bad	0.213	0.787	-0.574
Like	0.457	0.543	-0.086

According to Eq. 4.3, the word *"Pleasant"* from the previous example has the polarity score 0.46 $(2 * 0.73 - 1)$ and reflects that the word carries positive sentiment. Sentiment scores of some other words from the AUTO lexicon are presented in Table 4.7. We can observe that the words *"Good"* and *"Excellent"* have positive polarity scores well above zero, indicating the "positiveness" of these words. The word *"Bad"* has a negative polarity score weel below zero, as we would expect. The word *"Like"*, however, has the polarity score very close to zero, -0.086, indicating its neutrality. One might expect that the word *"Like"* should carry positive sentiment. To understand why the polarity score for the word *"Like"* is close to zero I investigated possible semantic roles of this word in English language:

(a) Being a verb to express preference. For example: *"I like ice-cream"*.

(b) Being a preposition for the purpose of comparison. For example: *"This town looks like Brighton."*

The first sentence has positive sentiment, however can easily be transformed into a negative sentence: *"I don't like ice-cream"*. This demonstrates that the word *"Like"* can be used with equal frequency for expressing positive and negative opinions. In the second example, the word *"Like"* is playing a role of a preposition and does not effect the overall polarity of the sentence. Thus, the word *"Like"* was correctly assigned a score close to zero using the approach described above.

In order to exclude words that do not help to classify the text as positive or negative, I removed all words from the AUTO lexicon with the polarity score in the range [-0.2; 0.2]. The final AUTO lexicon was comprised of 33,170 positive words and 19,429 negative words.

4.4.2 Lexicon Performance Results

To evaluate the impact of emoticons, abbreviations and slang on sentiment classification, as well as the quality of automatically generated lexicon AUTO, I used OL lexicon as the baseline and created two combinations of lexicons as shown in Table 4.8.

Table 4.8: Combinations of lexicons tested in the lexicon approach.

Lexicons combinations	Description	Positive tokens N	Negative tokens N
OL	Classic Opinion Lexicon	2,006	4,783
OL + EMO	Opinion Lexicon enhanced with emoticons, abbreviations and slang words, commonly used in social media	2,248	4,945
OL + EMO + AUTO	Opinion Lexicon enhanced with emoticons, abbreviations, slang and additional words obtained through automatic lexicon generation process	35,418	24,374

For each of the lexicon combinations in Table 4.8 I calculated sentiment scores for all tweets in the test set following the algorithm described in section 2.1.3 and using Eq. 2.1. A schematic representation of the algorithm is presented in Fig. 4.3. If none of the words in a particular tweet appeared in the lexicon, the algorithm would treat such tweet as objective text and would assign it the sentiment value 0.

Since sentences in the training sets are labelled with categorical values, such as ″pos″, ″neg″ or ″neut″, it is necessary to assign positive, negative or neutral labels to the tweets in the test set based on the calculated sentiment scores. One approach to assign the labels would be setting up the thresholds, for example, sentences with the scores in the range [-1; -0.3] should be labelled as negative, sentences with the scores in the range [-0.3; 0.3] should be classified as neutral and sentences that have sentiment score in the range [0.3; 1] should be marked as positive. However, the choice of the threshold is subjective. As an alternative way in this study I run a k-means clustering algorithm (MacQueen, 1967) and clustered sentiment scores into 3 classes using the sentiment score of each sentence as a feature in the clustering algorithm. I then compared the predicted labels against the true labels of tweets and calculated the accuracy of classification using Eq. 2.9.

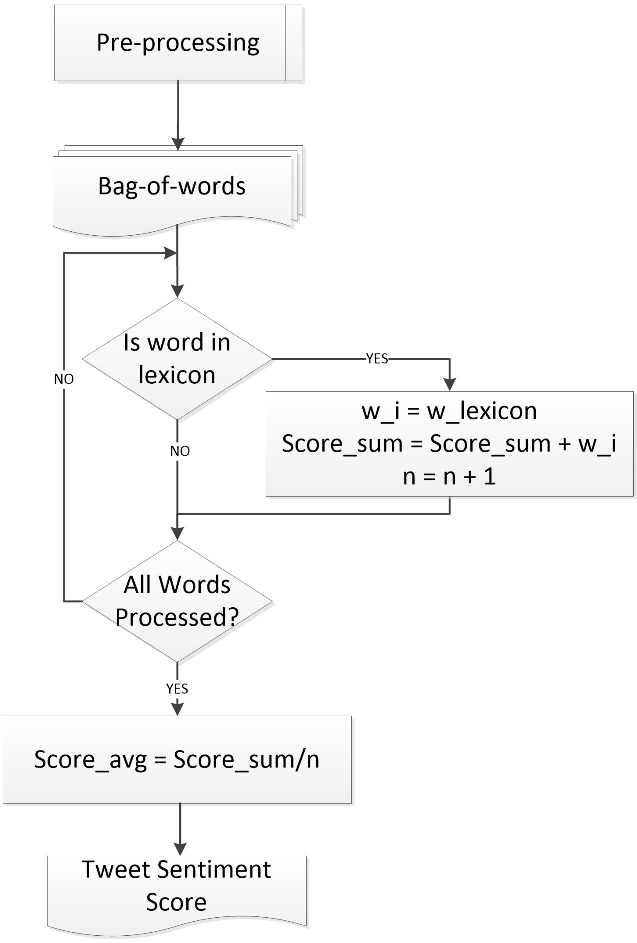


Figure 4.3: A schematic representation of the algorithm for the sentiment score calculation.

The results of K-means clustering for 2 types of test sets are presented below.

Results for the combined tweets and movies reviews dataset. Table 4.9 presents the results of sentiment classification for the 2-classes combined tweets and reviews dataset. We can see that when the OL lexicon was used, about the third of all messages in the test set (2,313 messages) were not classified as positive or negative. This means that the average sum of sentiment scores in non-classified messages was equal to zero or that none of the words from those messages appeared in the OL lexicon. The sentences which were not classified as ″positive″ or ″negative″ were assigned a ″neutral″ label, however this behaviour should be considered as a classification failure, since all the sentences in the test data set were carrying positive or negative sentiment.

While the overall accuracy of classification, calculated using Eq. 4.4, was only 46% (row ″Accuracy all″ in Table 4.9), the accuracy of distinguishing between

Table 4.9: Accuracy of 2-class classification for the combined tweets and movies reviews dataset using different lexicon combinations.

	OL	OL + EMO	OL + EMO + AUTO
Accuracy all	46%	48%	70%
Accuracy 2-classes	69%	70%	72%
Correct predictions N	3,260	3,364	4,953
Incorrect predictions N	1,478	1,461	1,935
Neutral N	2,313	2,226	163

positive and negative sentences was 69% (row "Accuracy 2-classes" Table 4.9), calculated by using Eq. 4.5.

$$\text{AccuracyAll} = \frac{\text{CorrectPredictions}}{\text{CorrectPredictions} + \text{IncorrectPredictions} + \text{Neutral}} \tag{4.4}$$

$$\text{Accuracy2Classes} = \frac{\text{CorrectPredictions}}{\text{CorrectPredictions} + \text{IncorrectPredictions}} \tag{4.5}$$

When the OL lexicon was enhanced by the EMO lexicon additional 87 messages were assigned sentiment labels and the accuracy increased slightly, to 48% and 70% for overall accuracy and for 2-classes accuracy, respectively. After the lexicon was enhanced further with the AUTO lexicon, almost all non-classified messages were assigned a positive or negative label. While the 2-classes accuracy improved only by 2%, the overall accuracy of classification improved significantly, by 24% compared to the OL lexicon only. This result indicates that lexicon enhancement leads to improved classification and that AUTO generated lexicon can be used for this purpose.

Results for the SemEval-2013 dataset. Sentiment analysis of the SemEval-2013 dataset is a more difficult task than the analysis of the combined tweets and movies reviews dataset, since it presents a 3-classes classification problem. Using Eq. 2.1, I calculated sentiment scores for test set based on three different lexicons OL, OL + EMO and OL + EMO + AUTO. Fig. 4.4 presents the histograms of scores for all three lexicons.

As it can be seen from the figure, the sentiment scores range between -1 and 1. The colors of the bars represent the true labels of the tweets: green stands for

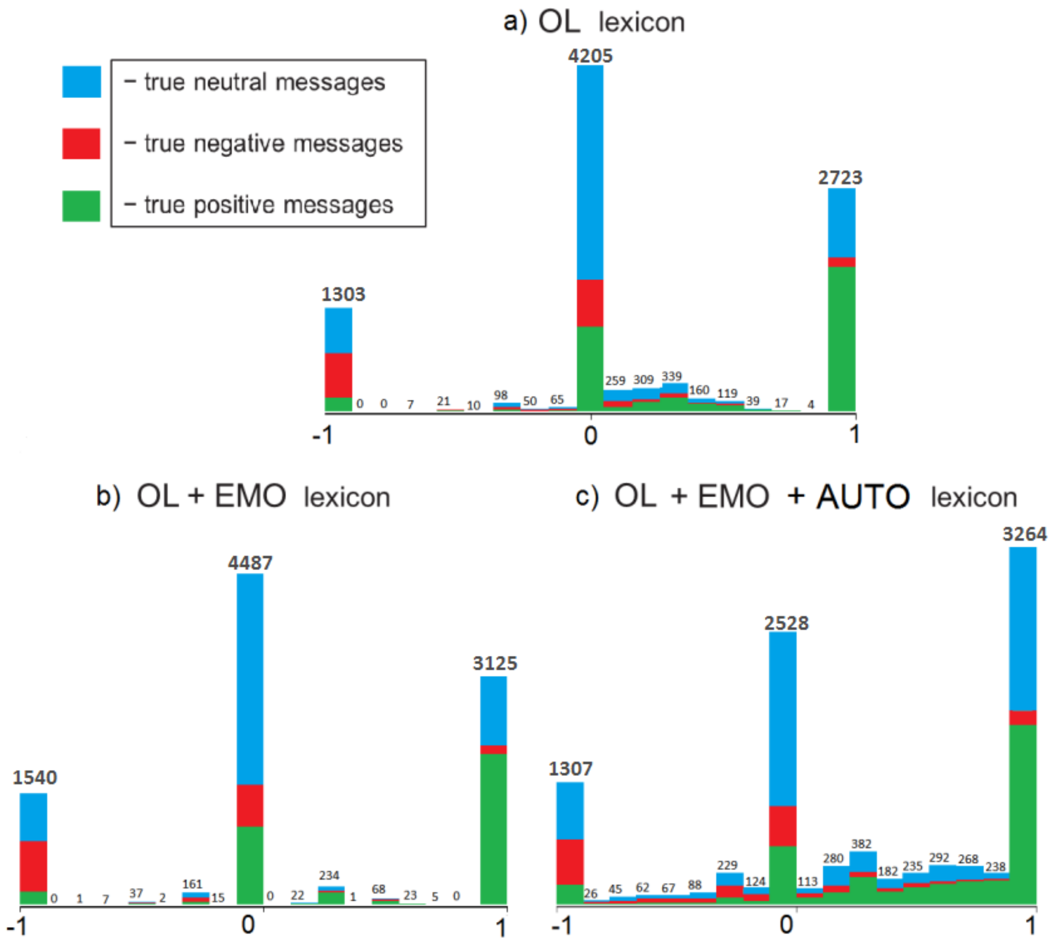


Figure 4.4: Histograms of sentiment scores for different lexicon combinations using the Simple Average Score. The colors of the bars represent the true labels of the tweets: green stands for positive messages, blue for neutral messages and red stands for positive messages.

positive messages, blue for neutral messages and red stands for negative messages. In the case of perfect classification, a clear separation of the colors would be obtained, with the green color bar on the right, close to the score ”+1”, representing positive messages, blue color in the middle around ”0”, representing neutral labels, and red color on the left, close to the score ”-1”, representing messages classified as negative. From Fig. 4.4 it can be seen that the biggest mistake for all types of lexicons was made when classifying neutral messages: the blue color is present for the sentiment scores of -1 and 1 in all three histograms, indicating that many neutral messages were classified as positive or negative. This could be explained by the fact that even neutral messages often contain one or more polarity words, which would lead to the final score of the message being different from 0 and being

classified as positive or negative. For example, the sentence:
"When the weather becomes good we are planing to go camping," does not express positive or negative opinion, however, it would be labelled as a positive sentence, since the word "good" appears in the lexicon as an indicator of the positive sentiment.

It is important to notice that when comparing the performance of different lexicons, the OL + EMO + AUTO lexicon had the least number of messages classified as neutral (2,528 vs. 4,205 for OL lexicon and 4,487 for Ol + EMO lexicon). Almost half of the messages that were assigned a score "+1" using OL + EMO + AUTO lexicon have blue color, indicating that their true label was "neutral". For the lexicons OL and OL + EMO the proportion of neutral messages classified as positive was much less - about one third of all messages that received a score "+1". These results suggest the AUTO lexicon contained words with assigned to them positive or negative scores that often appear in neutral sentences. This could happen because the training set from which the AUTO lexicon was generated was biased, or because the size of the training set was not large enough to allow correct calculation of the scores for all words. If a neutral word appeared more often in positive training sentences than in negative training sentences, that word could have a positive score assigned to it. If this word was then observed in a neutral sentence in the test set, the whole sentence could have been labelled as positive based on the positive score of that word. This means that one should aim to decrease bias or use large enough training sets for generating AUTO lexicons, especially when the classification problem involves neutral sentences.

As mentioned earlier, to assign a class label based on the calculated sentiment scores I run a K-means clustering. The accuracy of classification for three different lexicons after K-means clustering is presented in Table 4.10, row Score_Avg. I also proposed an alternative measure for the sentiment score, calculated as the logarithm of the average sum score and normalised in the range between [-1; 1], where -1 indicates the most negative score and 1 indicates the most positive score (Eq. 4.6). The purpose of the logarithm was to rescale the scores and allow a better separation between negative, neutral and positive classes.

$$\text{Score}_{\text{Log}} = \begin{cases} \text{sign}(\text{Score}_{\text{Avg}}) \log \mid 10\text{Score}_{\text{Avg}} \mid), & \text{if } \mid \text{Score}_{\text{Avg}} \mid > 0.1, \\ 0, & \text{otherwise} \end{cases} \tag{4.6}$$

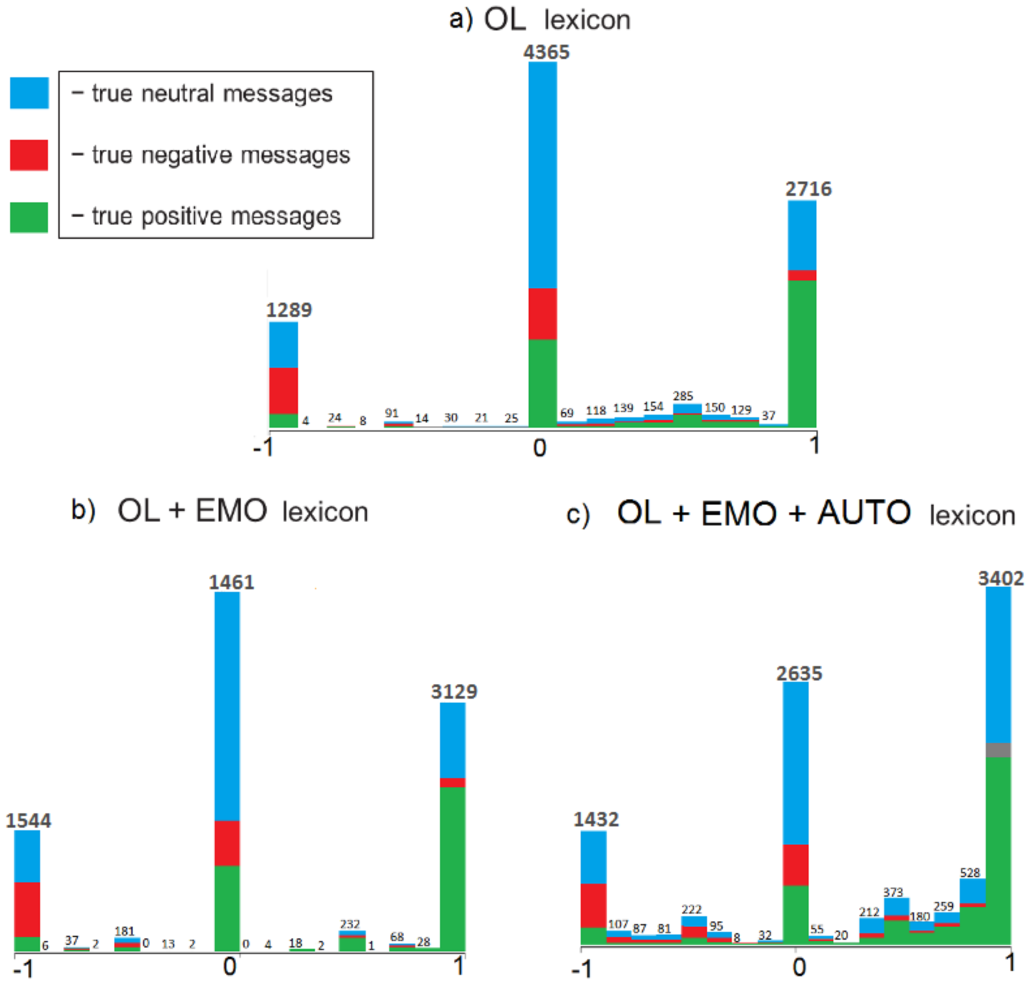


Figure 4.5: Histograms of sentiment scores for different lexicon combinations using the Logarithmic Score. The colors of the bars represent the true labels of the tweets: green stands for positive messages, blue for neutral messages and red stands for positive messages.

The histogram of sentiment scores in Fig. 4.5, calculated using the logarithmic approach, demonstrates that scores for positive, negative and neutral classes became more spread. The accuracy of sentiment classification after K-means clustering based on the logarithmic score is presented in Table 4.10, row Score_Log. We can see that the accuracy based on the logarithmic score is higher for all three lexicons when compared to the average sum score, demonstrating the ability of the logarithmic scale to provide a better separation of classes.

Table 4.10: Accuracy of classification for SemEval-2013 dataset for different lexicon combinations.

Accuracy	OL	OL + EMO	OL + EMO + AUTO
Score_Avg	57.07%	60.12%	51.33%
Score_Log	58.43%	61.74%	52.38%

When comparing the performance of three lexicons against each other, we can see that for 3-classes classification the OL + EMO lexicon with the logarithmic score outperformed OL and OL + EMO + AUTO lexicons. The outperformance over OL lexicon was expected and confirmed that enhancing the lexicon with emoticons, abbreviations and slang words increases the accuracy of classification. The decrease in performance when OL + EMO lexicon was enhanced with AUTO lexicon suggests that the quality of the training set, from which the AUTO lexicon is generated, has direct impact on final results. If the training set is biased or is not very large, the AUTO lexicon might contain words' scores that are not precise and, thus, decrease the accuracy of classification.

4.5 Sentiment Analysis using Machine Learning Approach

The machine learning approach is based on using a set of training records for training a classification model, where each record, also called an instance, has a class label assigned to it. The classification model is able to capture the properties, also called features, of the training records. A group of features related to one training record, is called a feature vector. Then, for a new instance of unknown class, the model can predict a class label for it.

4.5.1 Feature Generation.

In this section I describe the set of features that were constructed for the purpose of training a classifier.

- **n-grams presence.** During this step sets of uni-grams (single words) and bi-grams (two words) were created from the of consecutive words in each Twitter message. I created a binary feature that indicated the presence/absence of a particular ngram. Using "n-grams frequency" would not be logical in the case of Twitter text, since Twitter messages are very short, and a term is unlikely to appear in the same message more than once.

- **Lexicon score.** This feature is a sentiment score obtained during the lexicon based sentiment analysis, as described in the previous section using Eq. 4.6 and OL + EMO lexicon.
- **Elongated words number.** This feature represents the number of words with one character repeated more than 2 times, e.g. 'soooo'.
- **Emoticons presence.** A binary feature that indicates the presence/absence of positive and negative emoticons at any position in the tweet.
- **Negations number.** A numeric feature that represents the number of negated tokens.
- **POS numbers.** Numeric features that represent the number of occurrences for each part-of-speech in the tweet: verbs (posV), nouns (posN), adverbs (posAV), adjectives (posA), interjections (posI), etc.
- **Punctuation marks number.** A numeric feature that represents the number of occurrences of punctuation marks in a tweet.
- **Emoticons numbers.** A numeric feature that represents the number of positive and the number of negative emoticons in the tweet.
- **Sentiment words numbers.** These features represent the total count of tokens in the tweet that appear in the OL + EMO lexicon and have a score below and above 0;
- **Max and min scores.** These features represent the max and the min scores of tokens in the tweet that appear in the OL + EMO lexicon.
- **Last token score.** This feature is the lexicon score of the last token in the tweet. If the last token does not appear in the OL + EMO lexicon, the score is considered to be zero.

As it can be observed from the list of features above, many features are based on the sentiment scores of the tokens that appear in the OL + EMO lexicon. The list of features also includes a lexicon sentiment score, computed during the lexicon analysis step, as one of the features. This incorporation of information from the OL + EMO lexicon makes the proposed method to be a hybrid of the lexicon and machine learning approaches.

4.5.2 Feature Selection

In the machine learning classification task, the feature selection process is essential for improving classification accuracy and also providing greater insight into important class attributes (as described in section 2.1.4). In this study I used Information Gain evaluation algorithm with Shannon entropy (Shannon, 1948) measure, as in (Abbasi et al., 2008), and a Ranker search method (Ladha and Deepa, 2011b). Information Gain measures the decrease in entropy when the attribute A is present vs. when it is absent (see Eq. 2.4). The greater is the decrease in entropy when considering attribute A individually, the more significant feature A is for classification. A Ranker was used to sort features based on the information gain that they provide, from the most informative to the least informative. In the sentiment analysis only features for which the value of information gain was above zero were used. As the result, for the 3-classes benchmark dataset SemEval-2013 (Nakov et al., 2013), a subset of 528 features was selected from the total number of initial 1826 features. Top 50 selected features are displayed in Table 4.11.

Table 4.11: Top 50 features selected for SemEval-2013 dataset Nakov et al. (2013) based on the highest information gain.

1. LexiconScore
2. maxScore
3. posTokensNum
4. minScore
5. negTokensNum
6. good
7. love
8. pos_E
9. pos_R
10. pos_N
11. great
12. pos_V
13. happy
14. wait_neg
15. excited
16. can't
17. i
18. not
19. pos_A
20. pos_ElongWords
21. fun
22. lastTokenScore
23. i love
24. don
25. don't
26. amazing
27. f*ck
28. love you
29. can
30. awesome
31. bad
32. hope
33. thanks
34. luck
35. see you
36. looking forward
37. glad
38. wait_neg see_neg
39. thank
40. good luck
41. sad
42. nice
43. better
44. sorry
45. tomorrow_neg
46. so excited
47. proud
48. even_neg
49. congrats
50. sh*t

From Table 4.11 we can see that the "LexiconScore" feature, calculated using Eq. 4.6, appeared on the top of the list providing the largest information gain of all features. We can also see that many engineered features appeared among the most important features, such as "maxScore" (a lexicon score of the most positive word in a tweet), “posTokensNum” (a number of positive words in a tweet), “minScore” (a lexicon score of the most negative word in a tweet), “pos_E” (a number of emoticons in a tweet), “pos_R”(a number of adverbs in a tweet), “pos_N” (a number of nouns in a tweet), among others. Most of the top features were engineered features or unigrams, however, we also observe some bigrams, for example “i love”, “love you”, “see you”, “looking forward”, “good luck”, etc.

4.5.3 Classification Models

Each of the sentences from the training set was expressed in terms of its attributes. As the result, *n* by *m* binary matrix was created, where *n* is the number of training instances and *m* is the number of features. This matrix was used for training different classifiers:

1. *Decision Trees* (Mitchell, 1996) (section 2.1.5.1). A decision tree text classifier is a tree in which non-leaf nodes represent a conditional test on a feature, branches denote the outcomes of the test, and leafs represent class labels.

2. *Naive Bayes* (Narayanan et al., 2013) (section 2.1.5.2). In case of the sentiment analysis Naive Bayes classifier will test three hypotheses: one for each sentiment class. The hypothesis that has the highest probability is selected as a class of the text.

3. *Support Vector Machines (SVMs)* (Cortes and Vapnik, 1995, Vapnik, 1995) (section 2.1.5.3). For the linearly separable two-class data, the basic idea is to find a hyperplane, that not only separates the documents into classes, but for which the Euclidian distance to the closest training example, or margin, is as large as possible. In a three-class sentiment classification scenario, there are three pair-wise classifications: positive-negative, negative-neutral, positive-neutral.

4.5.4 Machine Learning Performance Results

Results for the combined tweets and movies reviews dataset. To validate the importance of the “Lexicon Sentiment” feature and other manually constructed features, I performed cross-validation tests on the combined tweets and movies reviews dataset according to two scenarios: i) in the first scenario (Table 4.12), I trained three different classifiers using only n-grams as features; ii) in the second scenario (Table 4.13), I trained the models using traditional n-grams features in combination with the "LexiconScore" feature and other manually constructed features: number of different parts-of-speech, number of emoticons, number of elongated words, etc.

Table 4.12: Scenario 1: 5-fold cross-validation test on a combined tweets and movies reviews dataset using only n-grams as features.

Method	Tokens Type	Accuracy	Precision	Recall	F-Score
Naive Bayes	uni/bigrams	81.5%	83.1%	81.2%	82.1%
Decision Trees	uni/bigrams	80.6%	81.0%	81.0%	81.0%
SVM	uni/bigrams	86.6%	86.8%	87.2%	87.0%

Table 4.13: Scenario 2: 5-fold cross-validation test on a combined tweets and movies reviews dataset using using traditional n-grams features in combination with manually constructed features: lexicon sentiment score, number of different parts-of-speech, number of emoticons, number of elongated words, etc.

Method	Tokens Type	Accuracy	Precision	Recall	F-Score
Naive Bayes	uni/bigrams	88.5%	88.6%	83.9%	86.2%
Decision Trees	uni/bigrams	89.9%	91.9%	88.2%	90.0%
SVM	uni/bigrams	91.2%	90.6%	91.5%	91.1%

As it can be observed from tables 4.12 and 4.13, the addition of the "Lexicon Sentiment" feature and other manually constructed features allowed to increase all performance measures for 3 classifies. For example, the accuracy of Naive Bayes classifier was increased by 7%, accuracy of Decision Trees was increased by over 9%, and the accuracy of SVM improved by 4.6%.

Results for the SemEval-2013 dataset. In this test I used a training set from SemEval-2013 competition, Task 2-B to train Naive Bayes, Decision Trees and SVM classifiers. I used features that performed best in the previous test, including lexicon sentiment score, presence/absence of N-grams, number of different parts-of-speech, number of emoticons, number of elongated words, etc. It is important to notice that the training dataset was highly unbalanced with the majority of tweets having a neutral label. In order to account for this unbalance, I also trained a cost-sensitive SVM model (Ling and Sheng, 2007). Cost-Sensitive classifier allows to minimize the total cost of classification by putting a higher cost on a particular type of error (in this case, misclassifying positive and negative messages as neutral).

As the next step, I tested the models on the unseen before test set from the same competition. While the classification was performed for 3 classes (pos, neg, neutral), the evaluation metric used in the competition was F-score (Equation 2.13), calculated for positive and negative classes only. The results of classification for different classifiers are presented in Table 4.14. We can observe that the Decision Tree algorithm had the lowest F-score of 62%. The reason may lay in a big size of the tree needed to incorporate all of the features. Because of the tree size, the algorithm needs to traverse multiple nodes until it reaches the leaf and predicts the class of the instance. This long path increases the probability of mistakes and thus decreases the accuracy of the classifier. Naive Bayes and SVM produced better scores of 64% and 66%, respectively. The best model was a Cost-sensitive SVM that allowed to achieve the F-measure of 73%.

Table 4.14: F-score results of sentiment classification using Naive Bayes, Decision Trees, SVM and Cost-Sensitive SVM classifiers. The test was performed on a test dataset from SemEval Competition-2013, Task 2-B Nakov et al. (2013).

Classifier	Naive Bayes	Decision Trees	SVM	Cost-Sensitive SVM
F-score	64.0%	62.0%	66.0%	73.0%

The results were compared against the results of 44 teams that took part in the SemEval-2013 competition and their 149 submissions. The top performing teams and the F-scores they achieved are presented in Table 4.15.

After comparing the results in Table 4.14 with the results of the competition (Table 4.15), we can conclude that the proposed hybrid algorithm based on the cost-sensitive SVM would had produced the best results, scoring 4 points higher than the winner of the SemEval-2013 competition. This is an important result, providing evidence that accounting for the unbalance in the training dataset allows to greatly improve model performance.

4.6 Chapter Conclusions

In this chapter I described the implementation process of two main approaches for sentiment analysis, a lexicon based method and a machine learning method, and proposed a new hybrid approach that is based on using lexicon approach results as

Table 4.15: Teams that produced top F-score results in the SemEval Competition-2013, Task 2-B Nakov et al. (2013).

Team name	F-score
NRC-Canada	69.02%
GUMLTLT	65.27%
teragram	64.86%
AVAYA	
BOUNCE	63.53%
KLUE	63.06%
AMI and ERIC	62.55%
FBM	61.17%
SAIL	
AVAYA	60.84%
SAIL	60.14%
UT-DB	59.87%
FBK-irst	59.76%

an input in machine learning classification.

In the lexicon based approach I compared the performance of three lexicons: i) an Opinion lexicon (OL); ii) an Opinion lexicon enhanced with manually created corpus of emoticons, abbreviations and social media slang expressions (OL + EMO); iii) OL + EMO further enhanced with automatically generated lexicon (OL + EMO + AUTO). I evaluated the performance on two benchmark Twitter datasets, the combined tweets and movies reviews dataset and SemEval-2013 dataset, and demonstrated that the OL + EMO lexicon outperformed the standard OL lexicon in both cases. These results revealed the importance of incorporating expressive signals such as emoticons, abbreviations and social media slang phrases into lexicons for Twitter analysis. The results are consistent with the later findings of Asghar et al. (2017), who also showed the importance of integrating emoticons, modifiers and domain specific words into sentiment analysis.

The results of my experiment also demonstrated that the OL + EMO + AUTO lexicon outperformed the OL and OL + EMO lexicons for a 2-classes classification problem (positive and negative classes), however, it produced the worst accuracy for a 3-classes classification task that also included a neutral class. These means that one should be careful when generating lexicons automatically from the training set. If the training set is not large enough, the scores in AUTO generated lexicon might be biased.

In the hybrid approach I used a sentiment score obtained during the lexicon based classification, along with other lexicon related features, as inputs for training machine learning classifiers. The ranking of all features based on the information gain scores during the feature selection process revealed that the sentiment score feature appeared on the top of the list, confirming its relevance in sentiment classification. I also demonstrated that in case of highly unbalanced datasets the utilisation of cost-sensitive classifiers improves accuracy of class prediction, in the proposed method the improvement of the cost-sensitive SVM over the regular SVM was 7%.

The proposed hybrid classification method, tested on the benchmark SemEval-2013 Twitter dataset, allowed to achieve F-score of 73%, which is 4% higher than the F-score of the winner of semEval-2013 competition.

Chapter 5

Quantifying Predictive Power of Twitter Events

This chapter presents a a framework for Twitter events detection, differentiation and quantification of their significance for predicting spikes in sales. A novel approach for clustering Twitter events based on their shapes is proposed that demonstrates that some types of Twitter events have the ability to predict spikes in sales. This chapter also describes a method for automatic identification of the optimum event window, solving a task of window selection, which is a common problem in the event study field. The framework is tested on a large-scale dataset of 150 million Tweets and sales data of 75 brands. Background and literature review for this chapter are presented in section 2.3

5.1 Introduction

Often, an economic objective of the firm is to increase sales for its products. However, unexpected increases in sales might be catastrophic for a business. If a company is not able to deal quickly and efficiently with the increase in demand, the sudden spike in sales might lead to the loss of customers and a rapid decline in profits. Therefore, one of the ultimate goals for management and operations teams of any business is to capture the early signals of growing demand so they could take necessary actions for optimising the resources.

Outbursts of activity on social media, particularly Twitter, related to a brand, might be an indication of a growing interest toward brand's products. In this chapter, I investigated a few spikes in Twitter sentiment (the results are presented in Appendix A, Tables 8.1 - 8.5). The results revealed that multiple types of social media events can be identified, such as: a) external events with strong brand association; b) internal brand events; c) social chatter driven by the brand; d) social chatter driven by people (see Conclusions of Appendix A). This information about spikes in social media could be beneficial for predicting the probabilities of spikes in sales. However, not all spikes of activities on social networks and blogs are associated with purchasing intentions. It has been highlighted by the scholars that differentiating between different types of events can shed some light on the problem of identifying the most predictive events. Sprenger et al. (2014), Werner and Murray (2004), Thompson (1988) performed the distinction between different types of events based on the topic that is being discussed: initial public offerings, or earnings announcements, or stock splits. I believe, that in the context of social media it is important to distinguish different types of events not only based on the content of the discussion, but also to take into account how information spreads through the network. Studying the internal dynamics of information diffusion is a challenging task and requires understanding of the rules of human collective behaviour. As discussed in section 2.3.1, Sornette, Crane and Sano et. al employed techniques from the complex systems theory and were able to get insights into the nature of spikes in blogosphere activities (Sano et al., 2013), views of Youtube videos (Crane and Sornette, 2008), Amazon books sales (Sornette et al., 2004) by studying the growth and relaxation signatures of those spikes. In this chapter, I propose a framework for events detection and differentiation based on their signatures in order to meet two objectives:

1. Quantify the relationship between Twitter events and sales events. This objective would allows us to understand whether non-filtered Twitter signal contains useful information about sales.

2. Quantify the relationship between specific types of Twitter events and sales events. The goal of this objective is to understand whether different dynamics of Twitter behaviour (events with different types of growth and relaxation signatures) have different effect on future sales, and to identify the dynamics that have significant predictive power.

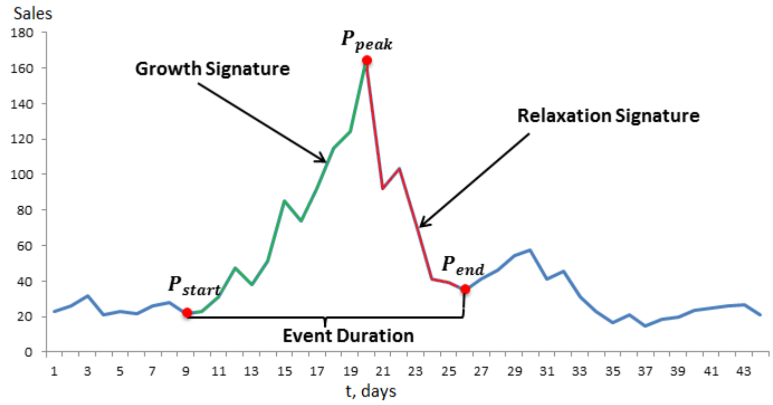


Figure 5.1: Signatures of a sales event. The green color represents the growth signature of the event, the red color represents the relaxation signature and the red circle is the peak of the event.

5.2 A Framework for Events Detection and Differentiation

To study spikes in sales and Twitter data I leverage the techniques from the event study field (Dolley, 1933). As outlined by MacKinlay (1997) to perform an event study it is necessary to identify the event date and finalise the event window within which the analysis is performed. In the social media era, however, defining an event date can be a complicated task. When critical information first emerges through Twitter, an event is not just a simple announcement moment, but a much broader phenomena which includes the dynamic of opinion propagation through the social network. In this context, I propose to define an event not by a single date, but to also consider other features: duration, the peak, growth and relaxation signatures. In this experiment the event is defined as follows:

Definition. An event is a quantitatively significant change of behaviour of a dynamic phenomenon over time. A Twitter or sales event is defined as an anomalous uplift of sentiment on Twitter or an extreme increase in volume of sales, respectively. Each event can be characterised by its duration (from few hours to few days), the growth signature, peak and the relaxation signature.

Fig. 5.1 presents an example of an event, where P_{start} denotes the start of the event, P_{peak} denotes the peak of the event, P_{end} represents the end date of the event, the subset of data points between P_{start} and P_{peak} is a growth signature, the subset between P_{peak} and P_{end} defines the relaxation signature.

In this chapter, I propose a framework for events detection, differentiation and

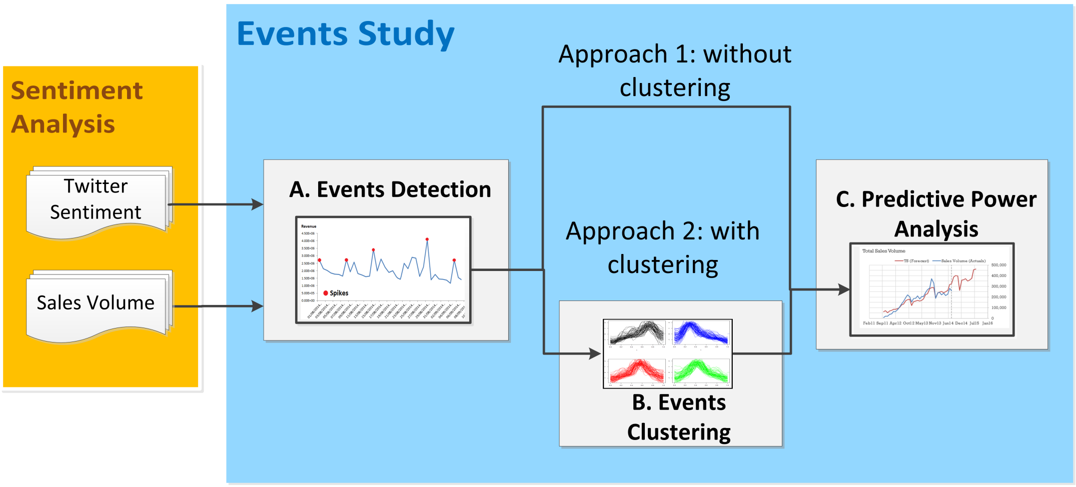


Figure 5.2: A framework for events detection and differentiation.

evaluation of whether events in one time series can be used to predict events in the other time series. The framework consists of three steps: (A) events detection; (B) events clustering; (C) predictive power analysis. To meet two objectives specified earlier I run the experiment according to two scenarios (schematic representation of the framework is shown in Fig. 5.2). In the first approach, I evaluate the power of the aggregated Twitter signal to predict sales events using steps A and D. In the second approach, I add an additional step B, during which a clustering of Twitter events based on their shapes is performed. I then calculate the predictive power of each Twitter cluster individually. The rational behind this process is to find categories of Twitter events that have a higher predictive power than the non-clustered Twitter signal.

Suggested framework can be used for analysis of any kind of time series, however, the necessary conditions for the time series under analysis are the following:

1. Time series should correspond to the same time period;
2. Time series should have the same aggregation window (1 hour, 1 day, 1 week, etc.).

In this chapter I demonstrate the application of the framework to retail brands using the following time series (created as described in section 3.5):

1. Daily Sales Volume time series, related to 75 brands over the period of one

year, from November 1, 2013, to October 31, 2014. Data was normalised using a z-score (Eq. 3.6.2)

2. Daily Tweets Sentiment time series that cover the same time period. The dataset includes sentiments of more than 150 million tweets that mention the names of selected brands.

In the subsections below I describe the different stages of the framework in more detail.

5.2.1 Step A: Events Detection

As stated in the definition, each event has a peak and growth and relaxation signatures. The process of event detection is a process of identifying the peaks and then extracting the growth and relaxation signatures.

Anomalous Peaks Detection.

The first step of the events detection process is identification of individual peaks in data that indicate anomalous behaviour. In the current experiment I compared the performance of three peaks (outlier) detection methods:

1. **ESD identifier** Rosner (1983) as described in section 2.3.2, Eq. 2.29
2. **Hampel identifier** Hampel (1971, 1974) as described in section 2.3.2, Eq. 2.31
3. **Median and InterQuartile Range (IQR)** Tukey (1977) as described in section 2.3.2, Eq. 2.33

When performing peaks detection it is important to consider that sales and Twitter time series are non-stationary. I observed the presence of weekly patterns in data, for example, tweets volume on Fridays and weekends was much higher than during the other days of the week. If Friday tweets volume was compared to Thursday tweets volume, the peak detection method would show a spike on Friday, however, I wanted to detect only special events and not regular bursts. Fig. 5.3 presents revenue data for one of the brands under analysis. Red circles in a graph indicate Fridays. As can be seen from Fig. 5.3 during most of the Fridays there is a peak in revenues. From the brand information I know that those peaks are related to weekly promotions that the company runs and the spikes in revenue on Fridays should not be considered special events. To account for non-stationarity, the three outlier detection measures described above (ESD/Hampel/IQR) were computed from the observations within a moving window that was comprised of data points from the same day of the week. For example, if the data point of primary interest was Friday, a moving window would include the data point of primary interest and a predefined number K of prior Friday values. In this way, Friday values were compared to the K previous Fridays only, Saturdays were compared to K previous Saturdays and so on.

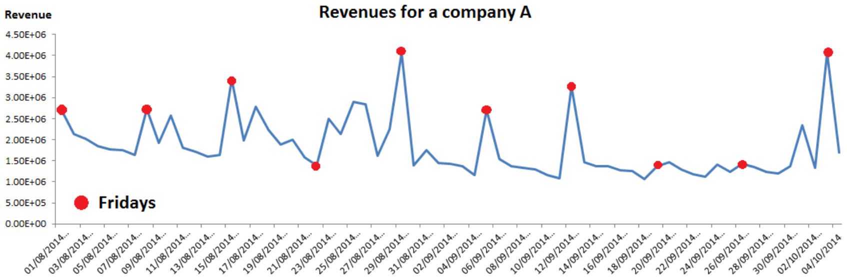


Figure 5.3: Daily revenues for a company A. The red circles indicate weekly Friday sales.

Extraction of Growth and Relaxation Signatures

For each peak P_{peak_i} that was detected during the first step, the goal is to identify the data point at which the event starts P_{start_i} and the data point at which the event finishes P_{end_i}.

Let us define the points at which the time series change its direction as change-points, and the time intervals between consecutive change-points as time segments. In Fig. 5.4 points $C = C_1, C_2, ..., C_9$ represent change-points. To extract the growth signature, the immediate left neighbouring point of P_{peak_i} is analysed to determine whether the point is a change-point. The procedure is repeated with consecutive left neighbours until the stopping criterion is met. The first change-point on the left side from P_{peak_i} that meets the stopping criterion is considered to be the start point of the event P_{start_i}. To extract the relaxation signature, identical procedure is performed with right neighbours of P_{peak_i}. The first change-point on the right side from P_{peak_i} that meets the stopping criterion is considered to be the end point of the event P_{end_i}.

The stopping criterion is considered to be met if any one of the following three conditions is fulfilled:

1. The first condition is fulfilled if the distance between the current change-point

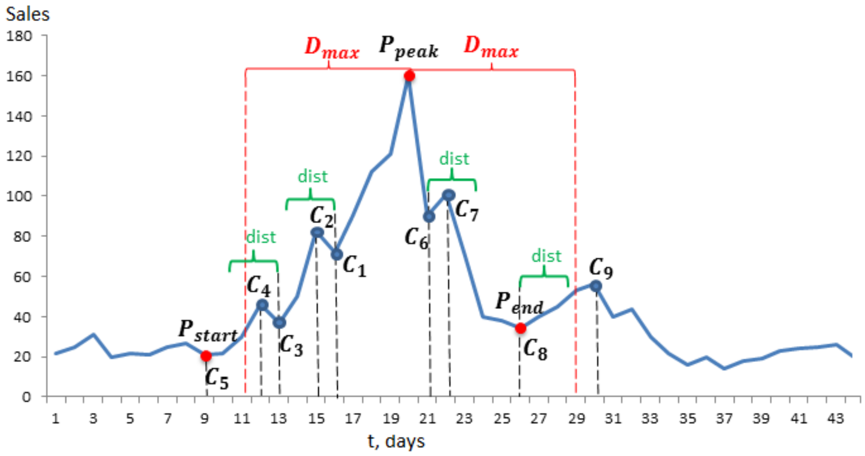


Figure 5.4: An example of a sales event. $C = C_1, C_2, ..., C_9$ denote change-points. D_{max} and $dist$ are two measures used in the first and second stopping conditions.

C_k and the peak P_{peak_i} exceeds the maximum distance D_{max}, predefined by the user. This condition allows to limit the duration of the event. Formally,

$$(C_k - P_{peak_i}) > D_{max} \tag{5.1}$$

2. The second condition is fulfilled if the distance between the current change-point C_k and the next change-point C_{k+1} exceeds the distance $dist$, predefined by the user. This rule allows to include points of changing direction, that do not, however, effect the overall trend, as part of the signature. For example, points C_2, C_4 and C_7 in Fig. 5.4. Formally,

$$(C_k - C_{k+1}) > dist \tag{5.2}$$

3. The third condition is fulfilled if the y value of the current change-point C_k (a sales figure or a sentiment value that corresponds to C_k) became lower than the local median (the median calculated over a moving window).

To illustrate, the change-point C_5 in Fig. 5.4 became a starting event point P_{start}, since it fulfilled the first condition $(C_5 - P_{peak}) > D_{max}$; the change-point C_8 became the end point P_{end}, since it satisfied the second condition $(C_8 - C_9) > dist$.

All time segments between P_{start_i} and P_{peak_i} denote a growth signature, while all time segments between P_{peak_i} and P_{end_i} represent a relaxation signature.

The algorithm for extracting the growth signature of the event:

```
currentPoint = $X_{peak}$ − 1
stoppingCriterions = FALSE
while (stoppingCriterions != TRUE){
  isChangePoint = CheckIfChangePoint(currentPoint);
  if (isChangePoint){
    nextChangePoint = FindNextChangePoint(currentChangePoint)
    stoppingCriterions = AnaliseStoppingCriterions()
    if (stoppingCriterions != TRUE)
      currentPoint = nextChangePoint − 1
  }
  else
    currentPoint = currentPoint − 1
}
$X_{start}$ = currentPoint
```

To extract extract the relaxation signature of the event, one should replace "-1" with "+1" in the above code, and X_{start} with X_{end}.

5.2.2 Step B: Events Clustering

One of the objectives of this experiment is to identify different types of Twitter Sentiment events based on their shapes. This is equivalent to a time series clustering problem, where members of the same cluster exhibit high similarity to each other, while members belonging to different clusters exhibit low similarity. In this experiment, I performed clustering of the events using a distance-based method (as described in 2.3.3), since spikes represent the sub-sequences of data which are quite short, excluding the possibility to find patterns that can be described by features. To calculate the distance between data points of different events I first smoothed time series using kernel density estimation (KDE) (Lopez-Novoa et al., 2015), and then used two known methods, Euclidean Distance (ED), as described in section 2.34, and Dynamic Time Warping Distance (DTW), as described in section **??**. To perform clustering I used hierarchical clustering algorithm because it does not require to specify the number of clusters a priory and it allows to use both, ED and DTW similarity measures (it has been shown that K-means might fail to cluster centres when using DTW (Niennattrakul and Ratanamahatana, 2007)). I also proposed a new method, Euclidean Slopes Distance (ESD), that is based on representing the original time series with the first derivatives (slopes) of selected data points and computing the Euclidean

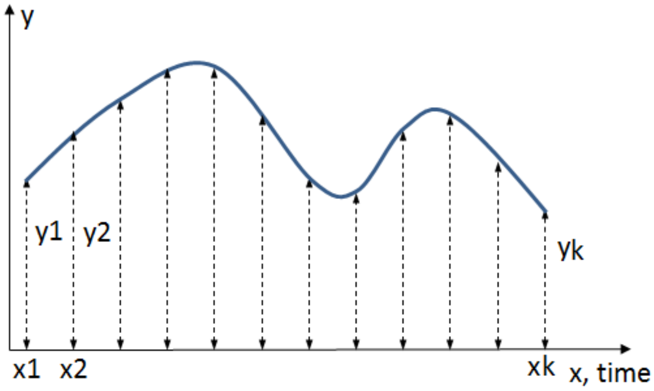


Figure 5.5: Schematic representation of time series divided into n stripes of equal width.

Distance between the slopes. For the data base of p time series $\{T_1, \ldots, T_p\}$, where $T_i = \{(x_i^1, y_i^1), ..., (x_i^n, y_i^n)\}$, the proposed algorithm works as follows:

- Each time series T_i is divided into K number of sequential stripes of equal width along the time-axis Toshniwal and Joshi (2005), where x_i is time and y_i is a corresponding value (see Fig. 5.5);
- Slopes S_i^k for each stripe of time series are calculated as

  $$S_i^k = \frac{(y_i^{k+1} - y_i^k)}{\Delta x} \tag{5.3}$$

  where (x_i^k, y_i^k) and (x_i^{k+1}, y_i^{k+1}) are the start end end coordinates for the k^{th} stripe of the i^{th} time sequence and Δx is the width of each of the stripes and is a constant;
- The Euclidean distance is computed between the corresponding stripes of time series;
- Hierarchical clustering is performed based on the distances obtained at the previous step.

The proposed approach works irrespective of the length of time sequences that are being compared, since it automatically normalizes all time series to equal length by dividing each time series into the same number of stripes K. The value of K should be optimally selected so it is neither too small (excessive computations) nor too large (loss of details). In this experiment parameter K was chosen to be equal to

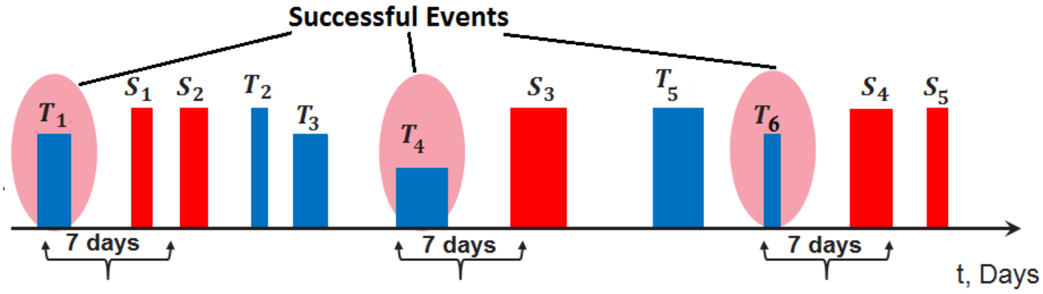


Figure 5.6: Schematic representation of successful Twitter events for the time window of 7 days. Red bars represent sales events, blue bars represent Twitter events, pink circles highlight successful Twitter events.

one third of the average length of the time series. I consider the proposed algorithm to be a combination of the distance and feature-based approaches since it operates on a reduced dimensionality approximation of the original time series, but still uses the Euclidean Distance as a metric for classification.

5.2.3 Step C: Predictive Power Analysis

Using Twitter sentiment time series and sales time series as a case-study, the null hypothesis, H0, can be defined as follows: sales events follow Twitter events in a random manner. To test the hypothesis I propose two significance tests.

5.2.3.1 Statistical Test One

This test evaluates how significant is the number of successful Twitter events. A Twitter event is considered to be successful if within a specified event window (7, 14, 21 days, etc.) there is a following it sales event (see Fig. 5.6). By defining the event window within which the analysis is performed, I follow a traditional approach of the event study (as proposed by MacKinlay (1997)).

To measure the significance of the observed number of successful events I randomise the positions of sales events 1000 times, calculate the number of successful events for each randomised scenario, and then compare the empirical results to the results after randomisation. The null hypothesis is rejected for a p-value of less than 0.05. If the number of observed successful Twitter events is outside of 95% confidence interval for the randomised case, I conclude that occurrence of Twitter events before sales events is not random.

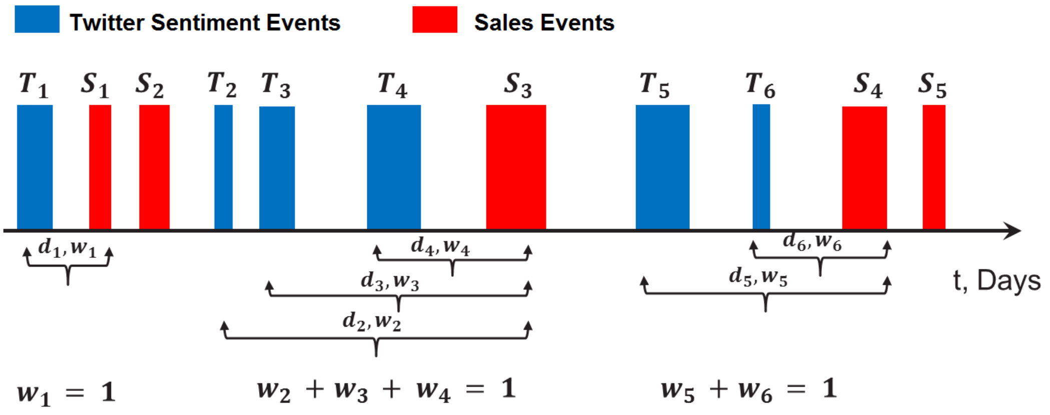


Figure 5.7: Schematic representation of sales events together with Twitter events and their weights. Red bars represent sales events, blue bars represent Twitter events.

5.2.3.2 Statistical Test Two

Selection of the event window, as performed in the first test, is a challenging task. In the second test, I propose an algorithm that allows to simultaneously measure the importance of events for all possible event windows and then automatically identify the windows at which Twitter events have a significant probability to be followed by sales events.

In this test, I consider that a Twitter event has a power to predict sales if at least one sales event appears after the beginning of that Twitter event. A sales event S_i might happen after a Twitter event T_i at different distances. For each Twitter event I store the distance d_i at which the first sales event happened. In the situations when multiple Twitter events are followed by one sales event I consider that all these Twitter events contributed to the appearance of a single sale event. I assign a weight w_i to each Twitter event, which is inversely proportional to the distance between the Twitter event and the following sales event: the longer is the distance between a sale and a Twitter event, the smaller is the weight, and vice a versa. The weights of Twitter events that have one sale following them sum up to one. This is the most conservative approach, which prevents me from over-counting the number of predictive events, however may result in under-counting them.

For example, in Fig. 5.7, I observe that a sale event S_3 has three Twitter events, T_2, T_3 and T_4, preceding it. I consider that all three Twitter events contributed to the appearance of the sales event S_3. Each of the Twitter events is being assigned a weight w_2, w_3, w_4, respectively, with the sum of the weights being equal to one: w_2

+ w_3 + w_4 = 1. The weights w_2, w_3, w_4 are inversely proportional to the distances d_2, d_3 and d_4 at which Twitter events occurred. For example, the longest distance is between the Twitter event T_2 and the sale event S_3, thus, T_2 event will have the smallest weight assigned to it. Conversely, the shortest distance is between event T_4 and the sales event, thus, it will be assigned the highest weight, assuming that Twitter event T_4 has the highest probability of contributing to the occurrence of the sale event S_3.

In this test, I am interested in analysing the probability of observing at least one sales event after a Twitter event for each event window. For this purpose I calculate a cumulative probability for each distance using the following steps:

- Calculate the time interval d_i between each Twitter event T and the first following it sales event S.

- Calculate the corresponding weight for each Twitter event w_i.

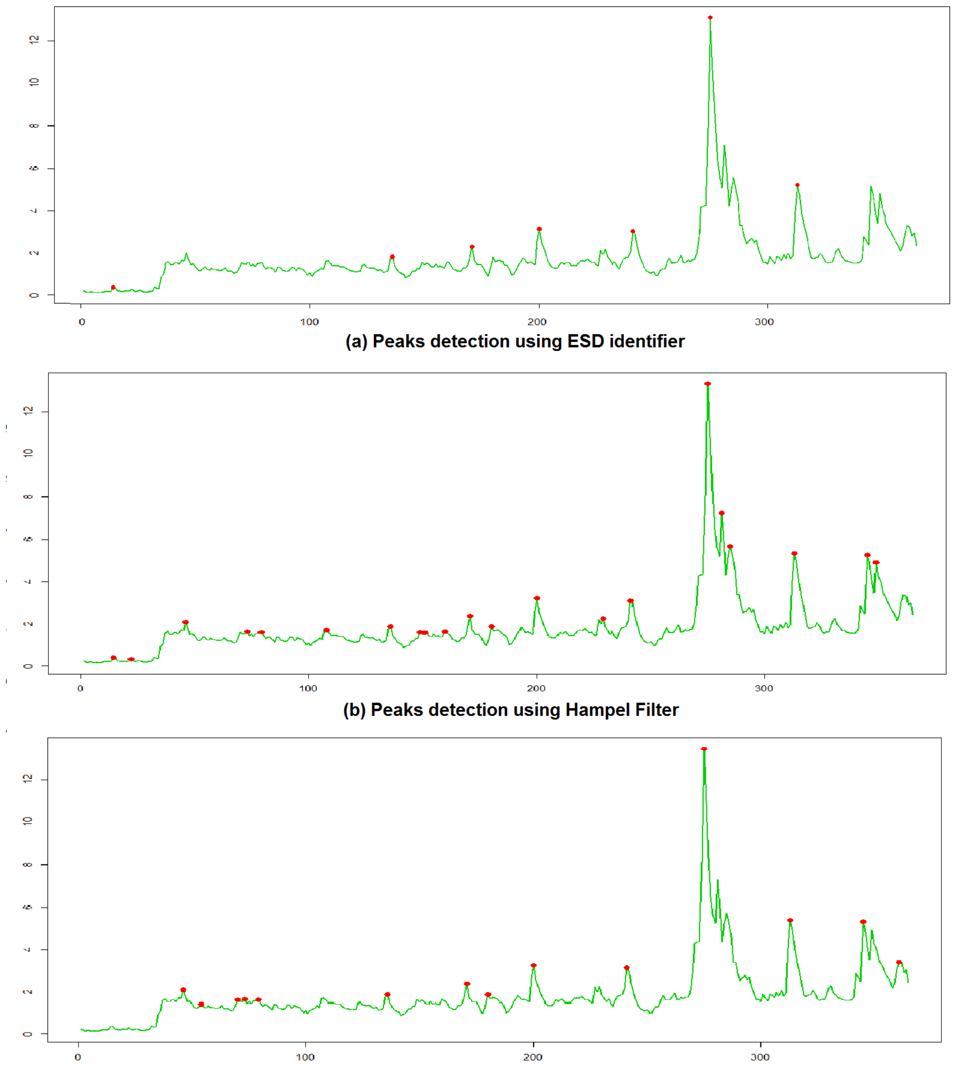
$$\begin{cases} w_i = \frac{w}{d_i} \\ \frac{w}{d_1} + \frac{w}{d_2} + \ldots + \frac{w}{d_n} = 1 \end{cases} \tag{5.4}$$

- Sort Twitter events in the incremental order of distances.

- For each Twitter event calculate the probability to have at least one sales event following it, by dividing the total number of events for each distance by the sum of their weights.

- Compute the cumulative probability for every distance by summing up the probabilities of the previous distances.

To identify the event windows at which Twitter events have significant power to predict sales events I perform a randomisation test. I randomise positions of all sales events, preserving their number and duration. The randomisation is made in a way that events do not overlap. The randomisation process is repeated 1000 times, and for each run the cumulative probability is calculated. To quantify the significance I compute the difference between the observed and randomised cumulative probabilities for each of 1000 runs. I then calculate the median difference and confidence intervals. I reject the null hypothesis if the 2.5



(a) Peaks detection using ESD identifier

(b) Peaks detection using Hampel Filter

(c) Peaks detection using IQR

Figure 5.8: Results of peak detection for sales data using three different methods: (a) ESD identifier; (b) Hampel filter; (c) IQR. Red dots denote peaks in time series.

percentile of the differences is higher than zero, which means that 97.5% of all differences are higher than zero. This allows us to conclude that that the observed system has a statistically significant advantage over the randomised system, and, therefore, sales are likely to follow Twitter events in a non-random manner.

5.3 Results and Discussion

5.3.1 Events Detection Results

The first step of events detection process is identifying abnormal peaks in time series. In the current experiment I used the moving window length of 7 days.

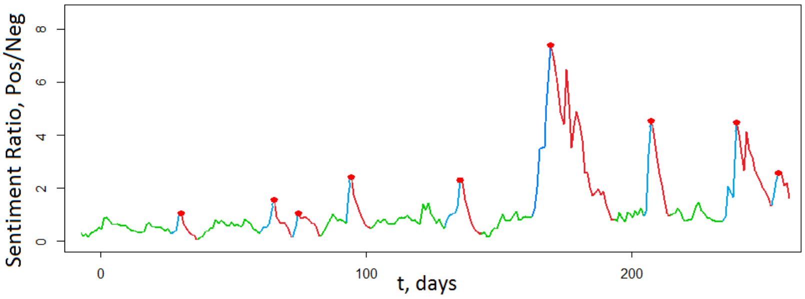


Figure 5.9: Signatures of Twitter sentiment events for a retail brand. The blue color represents the growth signature of the event, the red color represents the relaxation signature and the red circle is the peak of the event.

Fig. 5.8 presents the results of peaks detection for one of the brands using three approaches: (a) ESD identifier; (b) Hampel filter; (c) IQR. As it can be observed from Fig. 5.8, ESD identifier missed some of the spikes. The problem with this method is that often both the mean and the standard deviation are themselves extremely sensitive to the presence of outliers. In fact, if the level of outliers is higher than 10%, ESD detects no outliers at all. On the contrary, Hampel filter identified even small increases in sales as anomalous peaks. Hampel filter is much more resistant to the influence of outliers, however, it can be too aggressive in classifying values that are not really extreme as outliers. It can be shown that if more than 5% of the data points have the same value, MAD is computed to be 0, so any value different from the residual median is classified as an outlier. Comparing to ESD and Hampel filter, IQR method showed a superior performance. It captured the big spikes in data and did not suffer from identifying small increases in sales as outliers. I, therefore, used IQR as the method in the final analysis, since it provided a good balance between the amount of false-positives and false-negatives.

In the next step, I performed the extraction of events features (growth and relaxation) as described in section 5.2.1. Fig. 5.9 demonstrates the outcome of the event detection algorithm for one of the brands. The events of interest are highlighted with color: blue color represents the growth signature of the event, the red color represents the relaxation signature, and the red circle is the peak of the event.

As a result, 810 events in Twitter sentiment and 760 sales events were identified across 75 brands. I performed fitting of a power law ($y = a * x^b$) and did exponential fitting ($y = a * e^{bx}$) to growth and relaxation signatures of spikes after

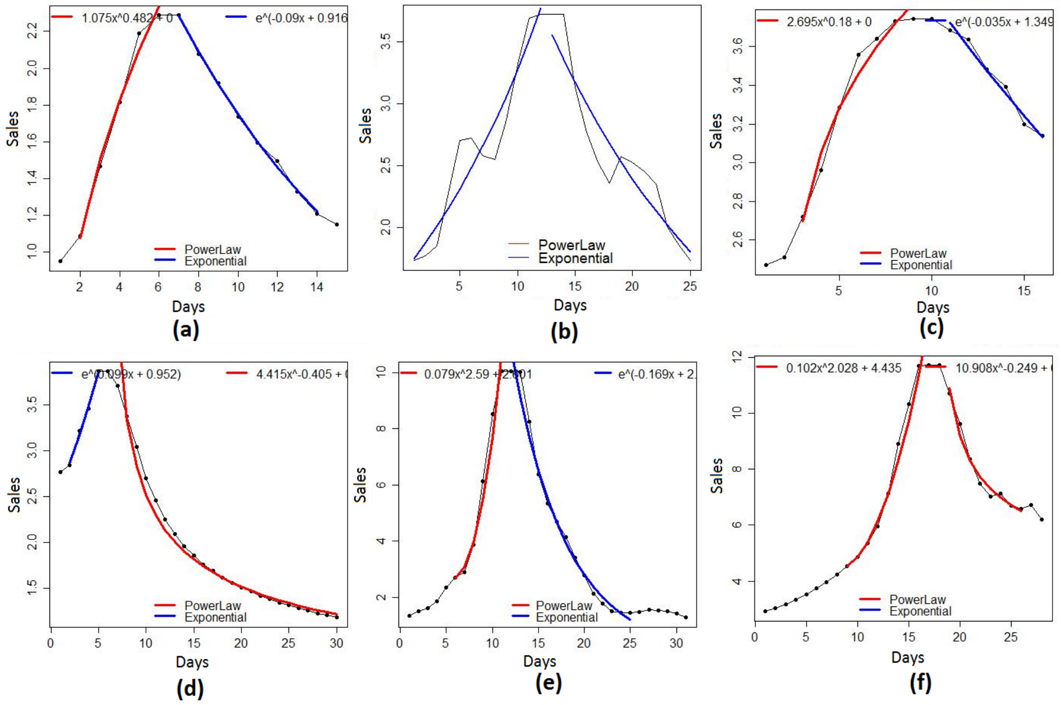


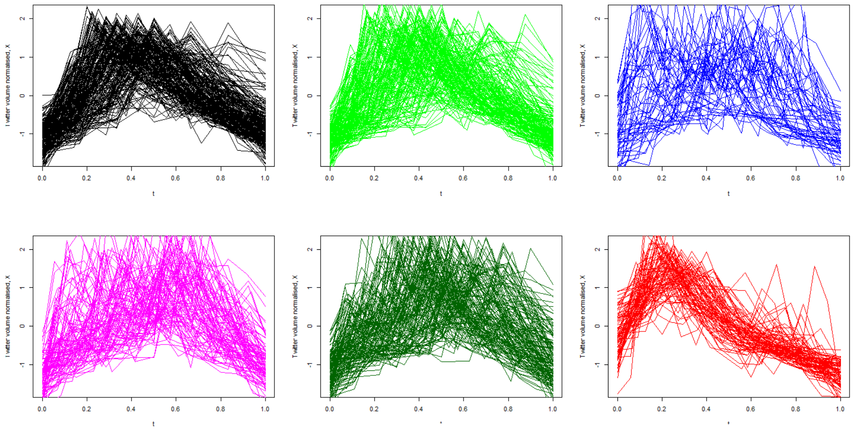
Figure 5.10: Examples of sales events with power law and exponential fitting to the growth and relaxation signatures. The red color means that power law is the best fit, the blue line means that the exponential fitting is the best.

KDE smoothing, and selected the best fit in order to observe different shapes. An example of power law and exponential fitting to some of the sales events is presented in Fig. 5.10. As it can be seen from the figure, different events clearly exhibit different dynamics of growth and relaxation.

5.3.2 Events Clustering Results

In this section I performed clustering of Twitter sentiment events based on their growth and relaxation signatures using hierarchical clustering with three different distance metrics ED, DTW, ESD as described in section 2.3.3. Clustering revealed interesting results: three distinct clusters of shapes were identified:

- A cluster with symmetric growth and relaxation dynamics (as in Fig. 5.10(a-c, e));
- A cluster with a short growth signature and a long relaxation signature (as in Fig. 5.10(d));
- A cluster with a long growth signature and a short relaxation signature (as in Fig. 5.10(f)).



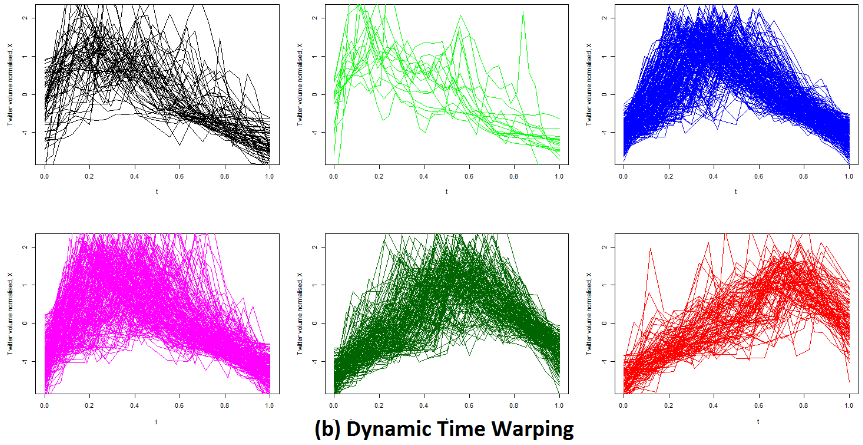
(a) Euclidean Distance

Figure 5.11: Twitter sentiment spikes clusters based on ED.

Based on the dendrogram I identified that the optimum number of clusters for Twitter sentiment was six, allowing to capture the three mentioned dynamics of events as well as variations in the widths of events. Clustering based on all three types of distance measures (ED, DTW, ESD) allowed to capture the three main dynamics of Twitter sentiment events, however, the measure of spread between cluster objects was different for the three measures:

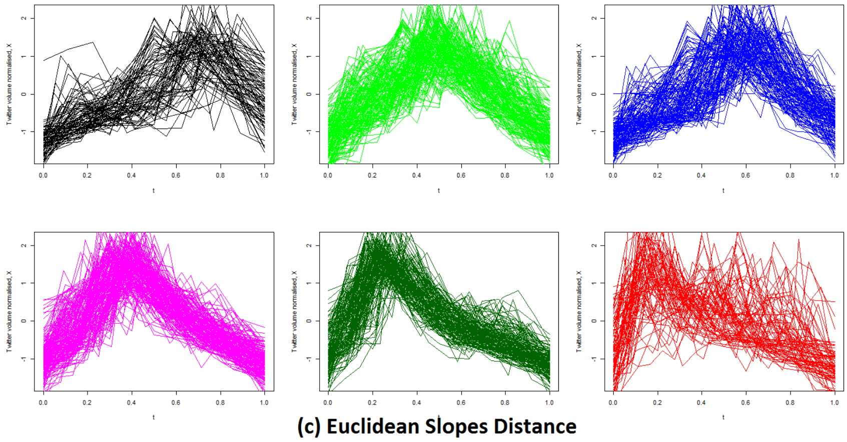
- **Euclidean distance results.** Extracted Twitter sentiment events had a duration in the range between 7 and 42 days. Events of different lengths might have growth and relaxation signatures of similar shapes and, therefore, should be clustered into one group. However, clustering based on ED failed to produce this outcome. Since ED performs linear mapping, in most cases the ED between the time series of similar lengths was smaller, than the distance between the time series of varying length. As the result, ED grouped the time series primarily based on their length (Fig. 5.11). Apart from that, Euclidean distance doesn't handle outliers, and it is very sensitive to signal transformations: shifting, amplitude and time scaling. These drawbacks made Euclidean inappropriate for current application.

- **DTW results.** DTW was designed to handle time sequences of varying length, solving the problem of ED. However, matching the shapes that do not line up in X-axis introduced a different kind of problem for the experiment at hand: DTW often grouped together time series only based on



(b) Dynamic Time Warping

Figure 5.12: Twitter sentiment spikes clusters based on DTW.



(c) Euclidean Slopes Distance

Figure 5.13: Twitter sentiment spikes clusters based on ESD.

the similarity of growth and relaxation signatures, ignoring the position of the peak of the spike in time (Fig. 5.12). Additionally, a non-linear mapping is computationally very expensive.

- **Euclidean slopes distance (ESD) results.** While Euclidean distance reflects similarity in time and dynamic time warping (DTW) reflects similarity in shape, the new ESD approach, based on the slopes of the stripes, was designed to capture both, similarity in time and in shape. Fig. 5.13 shows that the ESD method improved the quality of clustering results. The proposed approach has the following advantages:

  1. ESD is able to cluster time series of different lengths by automatically normalising them to an equal number of stripes. This resolves the

limitation of the Euclidean distance metric.

2. The higher level feature of the first derivative (slope) allows to extract information about the shape of time-series, thus, allows to capture the shape of the growth/relaxation signatures similarly to DTW.
3. ESD approach uses linear mapping between points which allows to capture the location of the peak in time.
4. By taking into consideration only slopes of stripes, the dimensionality of time series and, thus, noise, is reduced substantially.
5. Since the number of considered points is decreased, the ESD approach reduces the running time of clustering algorithm when compared to the other two metrics.

To further check the performance of the ESD clustering method I performed clustering of the benchmark time series datasets and compared the accuracies to the ED method. The results of clustering are presented in Appendix B, Figs. 8.1 - 8.4, and demonstrate that the proposed ESD method has equal or better performance than the ED. Since the Euclidean Slopes Distance approach outperformed the other two methods, I used event clusters obtained through ESD clustering for further analysis.

5.3.3 Scenario 1 Results: Predictive Power of the Non-Clustered Twitter Signal

In scenario 1, the analysis was performed before clustering Twitter events into different groups.

Table 5.1: Numbers of successful Twitter events for two event windows, 7 days and 21 days.

Sentiment events	Predictive within 7 days	Predictive within 21 days
Empirical Results	161	434
Randomised Results 95% CI	137 [135; 139]	418 [415; 420]

5.3.3.1 Statistical Test One

In the first significance test, I calculated the number of successful Twitter events for both, observed and randomised scenarios, as described in section 5.2.3.1. The analysis was performed for two event windows: 7 days and 21 days.

Table 5.1 presents empirical results along with the corresponding randomised sales intervals. For both event windows the number of empirically observed successful events appeared to be significantly larger than the number of successful events after randomisation of sales. It was empirically observed that 161 Twitter sentiment events had at least one sale event following them within 7 days. In case when the positions of sales events were randomised, the average number of Twitter events that had a sales event following them within 7 days was 137 with the 95% CI [135; 139]. We can see that the empirical result of 161 is outside of the 95% CI. Similar dynamic was observed for the event window of 21 days: 434 Twitter sentiment events were observed that had at least one sale event following them, while in case of randomization, the average number of Twitter events that had a sales event following them was 418 with the 95% CI [415; 420]. The empirical result of 434 is also outside of the 95% CI. The results for both, 7 days and 21 days window, suggest that the cases when Twitter events precede sales events do not happen by chance.

5.3.3.2 Statistical Test Two

In this test, I calculated a cumulative probability of having at least one sales event after a Twitter event for both, observed and randomised scenarios, as described in section 5.2.3.2.

In Fig. 5.14, the green line is the empirical cumulative probability and the black graphs are cumulative probabilities for the randomised events sequences. We can observe that the green graph is above the majority of the black graphs for about 20 days (approximately, 3 weeks). In order to quantify the difference between empirical and randomised results I calculated and plotted the median of differences between the observed cumulative probability and each of the probabilities after randomization, along with 2.5 and 97.5 percentiles (Fig. 5.15). If there was no underlying relationship between Twitter and sales events, the median of the differences between the observed cumulative probability and probabilities obtained after randomisation of sales events would be close to zero. However, Fig. 5.15 shows that the median is above zero along with 2.5 percentile of differences for the first 21 days. According to these results I reject the null hypothesis with the confidence of 97.5% and conclude that sales events follow

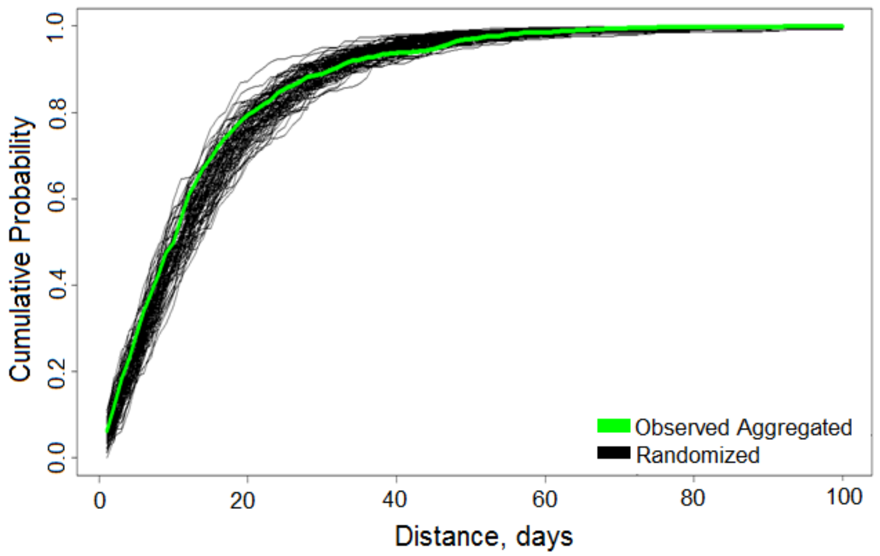


Figure 5.14: Cumulative probability of having a sale event within a specified time interval after a Twitter Sentiment event. The green line represents empirical results, the black lines represent randomised results.

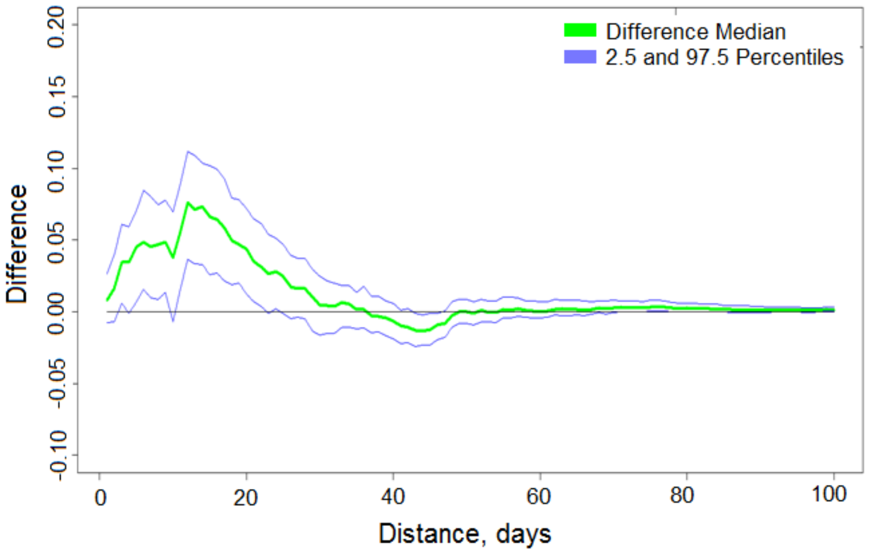


Figure 5.15: The median of differences between cumulative probabilities of observed and randomised data for Twitter Sentiment. The green line represents empirical results, the black lines represent randomised results and the blue lines represent 2.5 and 97.5 percentiles.

Twitter events in a non-random manner. The period during which the 2.5 percentile line is above zero defines the optimum event window. In this scenario it is 21 days, which means that there is a significant probability to observe a sale event after a Twitter event within 21 days after occurrence of a Twitter event.

5.3.4 Scenario 2 Results: Predictive Power of Different Twitter Event Clusters

In the second method I clustered Twitter events into six classes (Fig. 5.13) and analysed the predictive power of each Twitter class independently.

Statistical Test One

In this section, I tested the null hypothesis that Twitter sentiment events of different types appeared before sales events in a random manner. For this purpose I calculated the observed relative frequencies of successful Twitter sentiment events of different types (the notion of the successful Twitter event was explained in section 5.2.3.1). I then performed a randomisation test by randomising positions of sales events, identifying successful Twitter events within the same time horizon and calculating proportional representation of different types for the randomised sales. I repeated this randomisation test 1000 times and then calculated the average values with 95% confidence intervals (the third column of Table 5.2). I performed the analysis for the distances of 7 and 21 days after the Twitter event.

Table 5.2: Relative frequencies of successful Twitter spikes for success time horizon of 7 days.

Events	Observed	Random, 95% CI
Type 1	10.5	10.9 [10.4; 11.3]
Type 2	**23.6**	**20.2 [19.6; 20.9]**
Type 3	**13.7**	**16.9 [16.3; 17.5]**
Type 4	**26.1**	**22.2 [21.6; 23.1]**
Type 5	**16.8**	**20.2 [19.3; 20.7]**
Type 6	9.3	9.6 [9.2; 10.1]

Table 5.3: Relative frequencies of successful Twitter spikes for success time horizon of 21 days.

Events	Observed	Random, 95% CI
Type 1	11.1	10.9 [10.6; 11.1]
Type 2	20.5	20.3 [20.1; 20.5]
Type 3	**17.7**	**16.4 [16.2; 17.1]**
Type 4	22.4	22.6 [22.4; 23.1]
Type 5	**19.3**	**19.9 [19.4; 20.1]**
Type 6	**9.0**	**9.8 [9.6; 10.0]**

The results for a 7-days horizon are shown in Table 5.2, where the second column presents observed proportions of successful events of each type, and the third column presents proportions of each event type after randomisation along with 95%CI. From the observed 161 Twitter sentiment events, that were classified as successful within 7-days, the majority of events (26.1%) were classified as events of type 4, whereas the event type that had the smallest representation appeared to be class 6 (9.3%). Comparing the observed relative frequencies (the second column in Table 5.2) with the proportions after randomisation (the third column in Table 5.2), we can observe that the results for event types 2, 3, 4 and 5 are significantly different from random results. Specifically, event types 3 and 5 are significantly under-represented while event types 2 and 4 are significantly over-represented. Very similar results can be observed for the 434 events that were successful within 21 day (Table 5.3). We can see that the relative frequencies of types 3, 5 and 6 are significantly different from random.

Comparing the results for different distances, 7 and 21 days, we can observe that relative frequencies of events types change depending on the time interval. For example, the event of type 4 is over-represented for the 7 days distance and is not significant for the 21 days distance. This result is consistent with the analysis of cumulative probabilities (Fig. 5.17). Indeed, we observe that events of type 4 have significant predictive power during the first 1-2 weeks after the event and no predictive power on the day 21 and later.

Since we observed significant deviations from random in the success frequencies of different types of events, we can reject the null hypothesis and conclude that the occurrence of different types of Twitter events before sales events is not random. These means that different Twitter clusters have significantly different power to predict spikes in sales.

Statistical Test Two

As in the first scenario, I calculated the cumulative probability of having at least one sale event after the Twitter event, however, in this case the probability was computed individually for every Twitter event type. To compare the predictive power of each individual Twitter cluster I plotted the cumulative probability of the aggregated signal along with the results of individual event types (Fig. 5.16). I observed that for some event types their cumulative probability was higher than

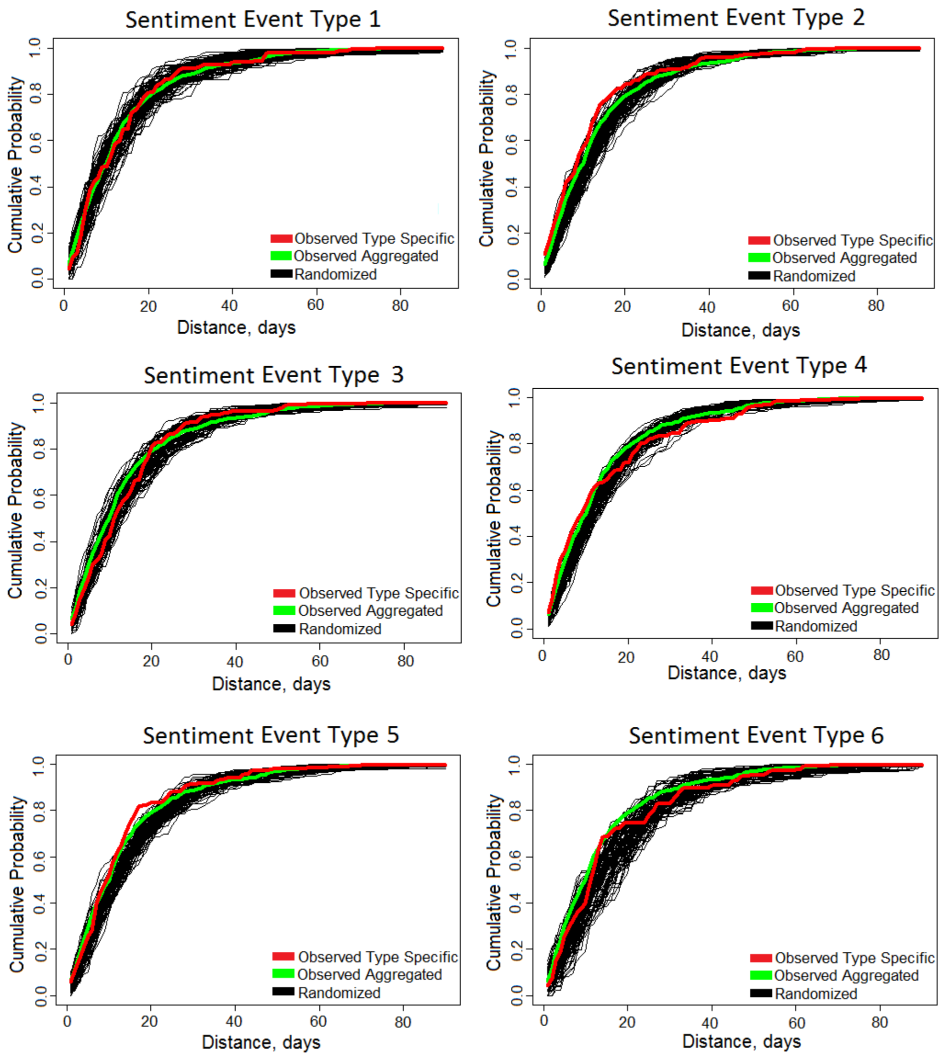


Figure 5.16: Cumulative probability to have a sale event within a specified time interval after Twitter Sentiment events of different types obtained using ESD clustering method. The red lines correspond to the cumulative probability of specific Twitter type, the green lines correspond to the cumulative probability of the aggregated Twitter signal, the black lines are the cumulative probabilities when sales positions are randomised.

the cumulative probability of the aggregated Twitter Sentiment signal for some periods of time. Calculating the median of differences between the empirical and randomised results along with 95%CI allowed to gain a better understanding of the dynamic (Fig. 5.17).

For events types 2, 4, 5 and 6 we can observe that: 1) in the first few days/weeks, the 2.5 percentile of the differences between observed and randomised results is above zero, which indicates that sales events follow Twitter

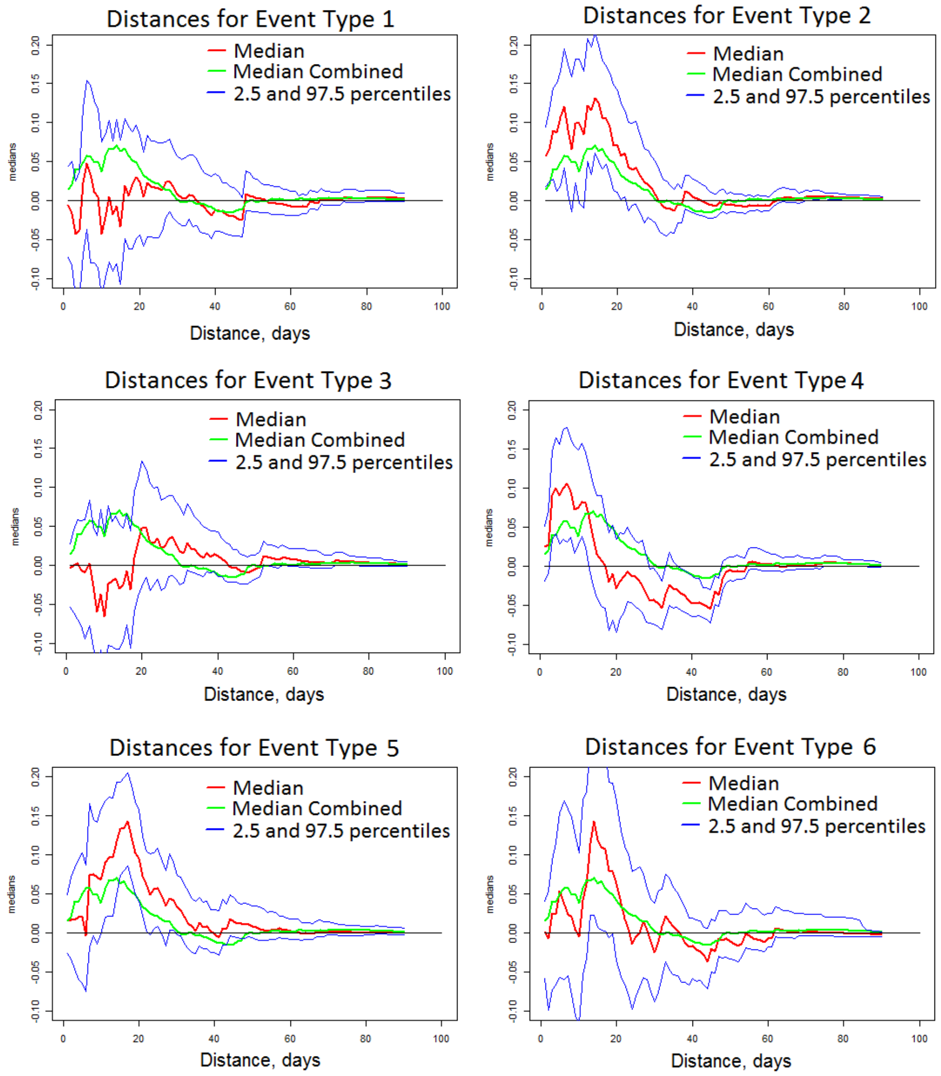


Figure 5.17: The median of differences between cumulative probabilities of observed and randomised data for Twitter Sentiment events of different types obtained using ESD clustering method. The red lines correspond to the median of differences for the specific Twitter type, the green lines correspond to the median of differences for the aggregated Twitter signal, the blue lines represent the 2.5 and 97.5 percentiles.

sentiment events of type 2, 4, 5 and 6 in a non-random manner; 2) the median of differences between observed cumulative probability for a specific Twitter type and the randomised sequences (red line) is greater than the median of differences between the observed cumulative probability for the aggregated Twitter signal and the randomised data (green line), which means that signals of the event types 2, 4, 5 and 6 have better predictive power than the aggregated Twitter signal. It is interesting to notice that event type 2 has a consistent significant predictive power during the first 3 weeks after the event (the 2.5 percentile is continuously above zero), while event type 4 shows significant predictive power only during the first two weeks; event type 5 is predictive between 12 and 22 days; event type 6 is only predictive for a short period of time during the third week after the event.

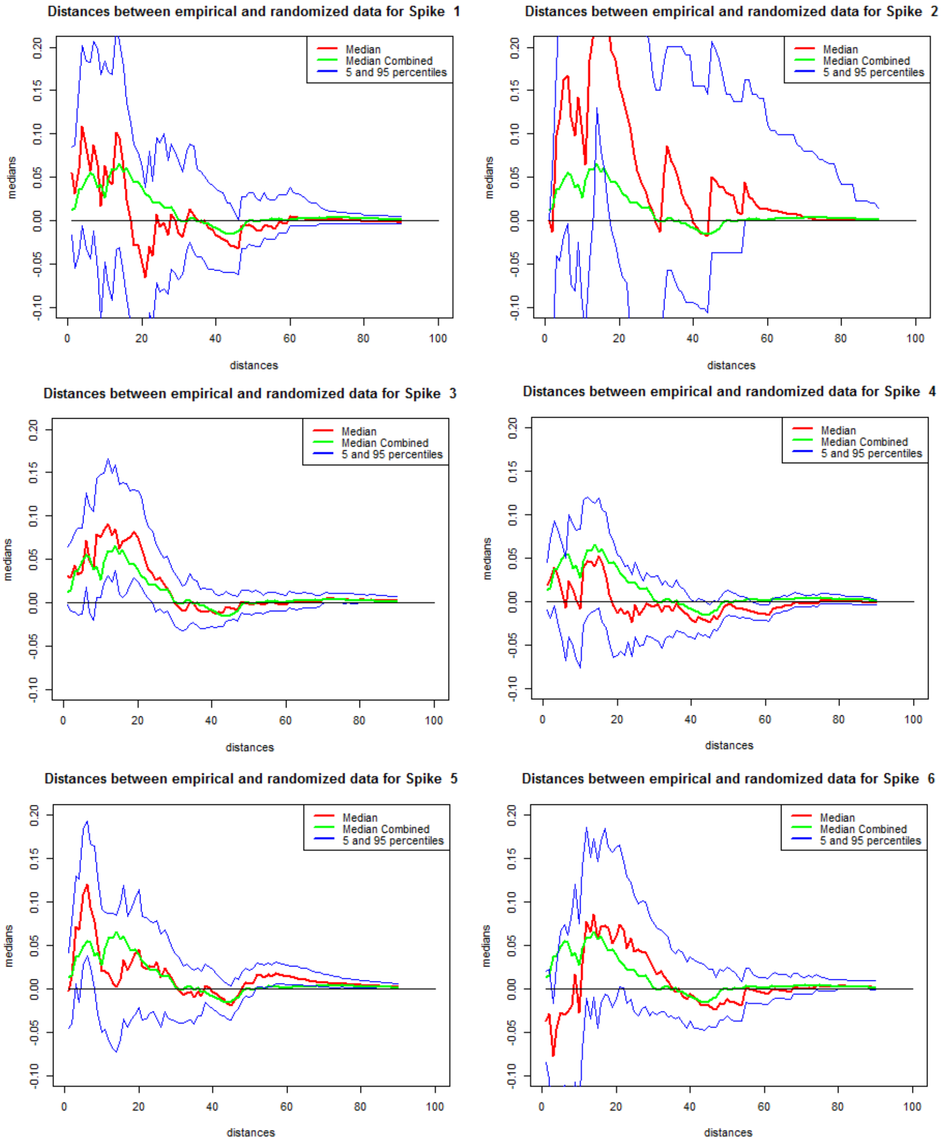


Figure 5.18: The median of differences between cumulative probabilities of observed and randomised data for Twitter Sentiment events of different types obtained using DTW clustering method. The red lines correspond to the median of differences for the specific Twitter type, the green lines correspond to the median of differences for the aggregated Twitter signal, the blue lines represent the 2.5 and 97.5 percentiles.

This information can be incorporated into forecasting models that consider the predictive power of different Twitter events at different distances. For example, the

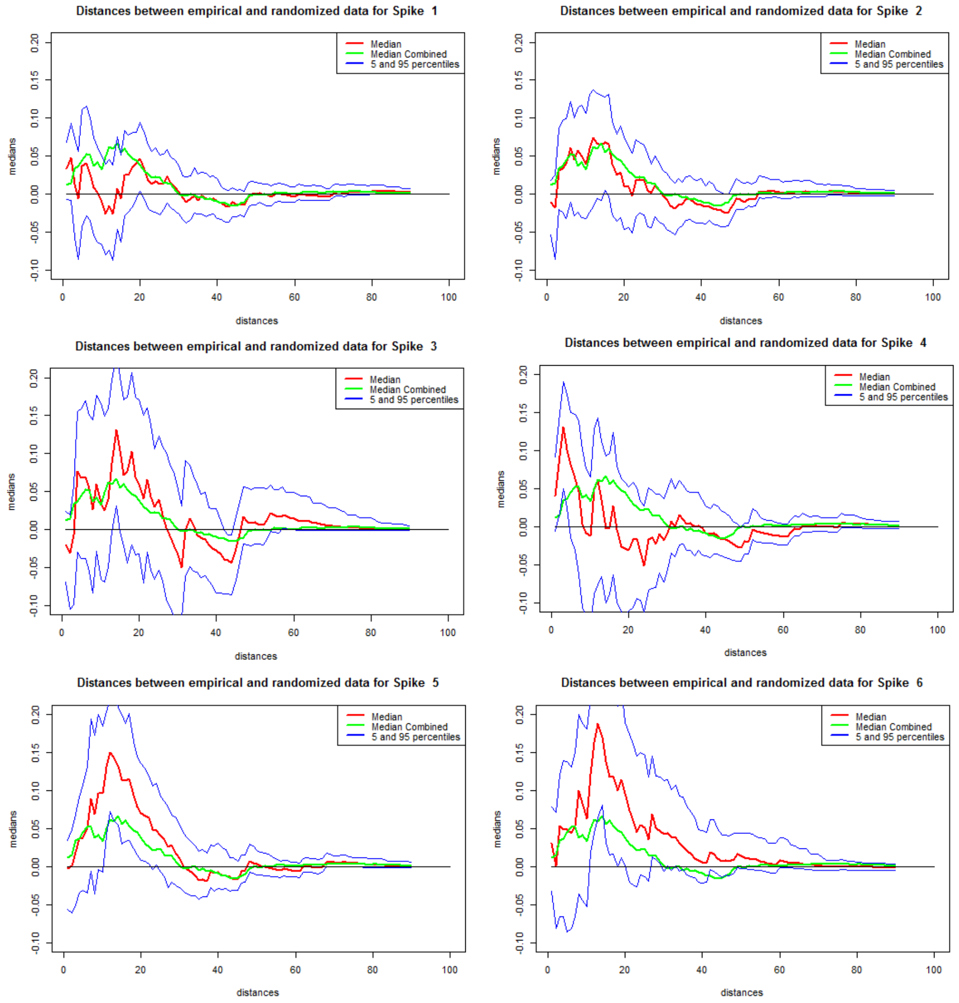


Figure 5.19: The median of differences between cumulative probabilities of observed and randomised data for Twitter Sentiment events of different types obtained using ED clustering method. The red lines correspond to the median of differences for the specific Twitter type, the green lines correspond to the median of differences for the aggregated Twitter signal, the blue lines represent the 2.5 and 97.5 percentiles.

Twitter Sentiment events of type 2 can be used to predict sales events within the first 3 weeks after the twitter event whereas Twitter events of type 4 can be used to predict sales events only within the first 2 weeks after a Twitter event.

I also performed the same statistical validation for the classes of events obtained through the other two clustering methods: based on the ED and based on DTW. The clusters obtained through DTW clustering (Fig. 5.18) did not reveal any significant predictive power. The results based on the Euclidean distance (Fig. 5.19) revealed some clusters of events that were followed by sales events in a non-random manner, however, the number of days for which the results were significant was less than for the results obtained through the ESD clustering. This analysis confirmed that the proposed method of slopes-based clustering was the most suitable for the problem.

5.4 Chapter Conclusions

In this chapter I studied the relationships between Twitter sentiment events and events in consumer sales. For this purpose I defined a framework for automatic events detection, events differentiation and evaluation of events importance. The framework was tested on a large dataset of sales for 75 brands and 150 million tweets. To the best of my knowledge, for the first time events analysis was conducted on a large scale for consumer products that do not necessarily receive large social attention.

This experiment presents a contribution to the field of event study. First of all, I proposed a new definition of the extreme event. I argued that in the current social media landscape it is important to define events not by a single date and an event window, but incorporate features such as a peak date, growth dynamic, relaxation dynamic and the duration of the event. Secondly, I proposed an approach that does not require to specify an event window a priory. The analysis of the predictive power of Twitter is performed along a wide range of windows and the statistical approach of calculating a cumulative probability allows to automatically identify the windows during which Twitter events have a significant probability to be followed by sales events. Thirdly, I suggested a novel method for events clustering based on the slopes of growth and relaxation signatures. The proposed method revealed that time series can be effectively approximated by higher level representations while still preserving their shape characteristics useful in classification or clustering tasks.

The predictive power of Twitter events in this experiment was evaluated using two scenarios: in the first scenario I performed the analysis of the aggregated Twitter signal considering only tweets' sentiment; in the second scenario I clustered Twitter sentiment events using the proposed approach and then calculated the statistics of successful predictions separately for every event type. The important result of the experiment is identification of types of Twitter events that have the power to predict events in sales and this predictive power is higher than the predictive power of the aggregated Twitter signal.

The findings of this chapter can be summarised as follows:

- There is a significant probability that Twitter sentiment events will be followed by spikes in sales within the next 3 weeks;
- Events can be clustered into categories based on their shapes (position of the peak, growth and relaxation signatures);
- Different Twitter event shapes are differently associated with sales;
- Some sentiment event types have significantly higher probability to be followed by sales events than events from the non-clustered Twitter signal.

Chapter 6

Application of Social-Media for Sales Forecasting

This chapter investigates whether information from social networks, such as Facebook and Twitter, and search data from Google Trends has the ability to improve daily sales forecasts. A methodology is proposed that allows to perform feature engineering, feature selection, stationarity and cointegration checks. Depending on the outcomes of stationarity and cointegration tests, the most appropriate model is chosen for the next-day sales forecasting. The models that are tested in this chapter include multiple linear regression, artificial neural networks and vector error correction model. Prediction of the next-day sales direction is also performed using logistic regression and artificial neural networks. The forecasting performance of different models is compared through a case study of 18 brands over the period of two years. Background and literature review for this chapter are presented in section 2.4

6.1 Introduction

Forecasting demand on the daily basis is crucial in the retail industry. It helps operational management to plan inventory management, place orders at various times of the day and reduce losses. Social media data such as Twitter messages, Facebook likes and Google Trends can be extremely useful for obtaining better sales forecasts. Although existing literature into the effects of Twitter (Souza et al., 2016, Dijkman et al., 2015, Antenucci et al., 2014, Culotta, 2013, Bermingham and Smeaton, 2011, Asur and Huberman, 2010), Facebook (Cui et al., 2017, Lee et al., 2016, Boldt et al., 2016, Du et al., 2015), blogs (Gruhl et al., 2005) and Google Trends

(Wu and Brynjolfsso, 2015, Preis et al., 2013, Choi and Varian, 2012, Kulkarni et al., 2012) on predicting different economic variables is vast, the literature is mostly limited to analysing one social media factor at a time. Also, previous research in sales forecasting has been mainly performed for products that receive a lot of social attention such as movies (Asur and Huberman, 2010) or books (Gruhl et al., 2005), failing to provide answers on the impact of social media for predicting sales for general retail products. The main objective of this chapter is to explore the effect of Twitter, Facebook and Google Trends, independently and jointly, on the sales of 18 companies across different sectors. The goal is also to take into account the non-stationarity of variables and possibility of non-linear relationships between them by training three types of models: multiple linear regression, artificial neural networks and vector error correction model.

To summarise, this chapter sets out to answer the following research questions:

1. To what extent information from social media helps to improve out-of-sample sales forecasts?

2. Which factors contain more information about future sales, Facebook, Twitter or Google Trends? Do they compliment each other in effecting sales?

3. Does Vector Error Correction Model allow to produce better forecasts by taking into account long-run relationship between sales and social media variables?

4. Does a non linear Artificial Neural Network model increase predictive power compared to linear models?

6.2 Methodology

The working hypothesis in this chapter is that social media information can be extremely valuable for short-term sales forecasts by reflecting changes in consumer demand within hours or days, therefore, the aim is to produce next-day sales forecasts. To check this hypothesis I suggest a framework as shown in Fig. 6.1.

During the first step of pre-processing I perform data cleaning, data transformation and normalisation (as described in 3.6.2), feature engineering and feature selection.

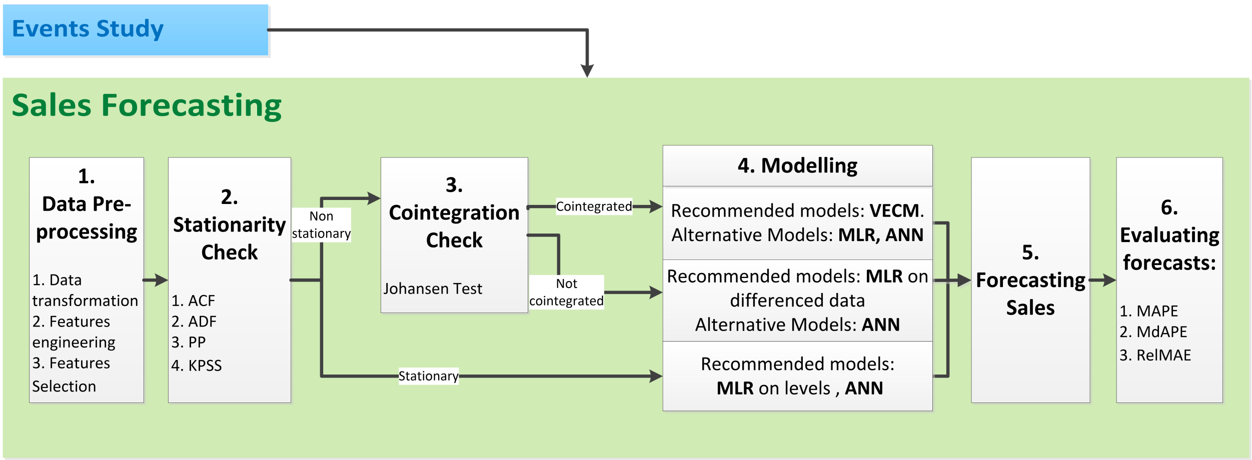


Figure 6.1: A framework for sales forecasting.

In order to avoid ″spurious regressions″ and making wrong conclusions about the relationships between sales and social media variables, in the second step I perform stationarity analysis (as described in 2.2.3) and determine the order of integration for each variable (Turner, 2000). The order of integration is the number of times that the series has to be differenced in order to make it stationary. If time series is stationary the order of integration is zero and is denoted as I(0). I use the notation I(n) to refer to a variable whose nth difference is stationary and the term ″levels″ to denote the actual values of time series. I perform the cointegration check during the third step of the analysis as described in 2.2.6.

Depending on the order of integration of each variable calculated during the stage 2 and the cointegrating relationship between the variables obtained at step 3, the most appropriate model is chosen following the guidance provided by Giles (2013):

1. All of the series are I(0), and hence stationary. In this case, modelling of data can be done at their levels, using Multiple Linear Regression estimated by Ordinary Least Squares (OLS).

2. All of the series are integrated of the same order, e.g. I(1), but they are not cointegrated. Granger and Newbold (1974) presented strong evidence that regressions involving random walks are spurious when performed on the levels, but not on the differences. Thus, in this scenario a standard regression model on the appropriately differenced time series can be used.

3. All of the series are integrated of the same order, and they are cointegrated. In this case, two types of models can be estimated: (i) An OLS regression model using the levels of the data. This will provide the short-run equilibrating relationship between the variables. (ii) A Vector error-correction model (VECM), estimated by OLS. This model will represent the long-run dynamics of the relationship between the variables.

4. The last case, which is outside the scope of this chapter, is when some of the variables are I(0), some may be I(1) or even fractionally integrated, and there is also the possibility of cointegration among some of the variables. In this situation an ARDL model proposed by Pesaran et al. (2001) might be appropriate.

According to the guidance above, during the stage 5 of the forecasting framework I run Multiple Linear Regression on levels for stationary variables, or on differenced data for non-stationary and not cointegrated variables; I use Vector Error Correction Model in case variables are non-stationary, but cointegrated. In this study, for the cases when VECM is recommended, apart from VECM I also run MLR in order to answer the question 3 mentioned at the beginning of the chapter. Also, for all types of time series I run an Artificial Neural Network in order to compare its performance to the performance of linear models and answer the question 4. Apart from regression models described above, that are used to predict the volume of next day sales, I also run classification models to predict the direction of the next day sales: Logistic Regression (LR) and ANN.

The performance of models is evaluated at stage 6. Since the number of observations for each brand is limited (only 2 years' worth of data), I used the time series cross validation technique described in section 2.4.3.1. To compare the forecasts of the out-of-sample dataset against a benchmark I used metrics described in section 2.4.3.2: Mean Absolute Percentage Error (MAPE), Median Absolute Percentage Error (MdAPE), Relative Mean Absolute Error (RelMAE) and performed the Diebold-Mariano test (Diebold and Mariano, 1995) (DM-test).

6.2.1 Datasets

The time series used for empirical investigation in this chapter were the following (created as described in section 3.5):

1. *Sales volume.* Daily globally aggregated sales for 18 brands for the period of two years, from November 1, 2013, to October 31, 2015.

2. *Twitter.* Two kinds of time series were used: (i) Tweets daily volume time series; (ii) Tweets daily sentiment time series.

3. *Facebook.* Three kinds of time series were used based on Facebook data: (i) Facebook daily volume of posts; (ii) Facebook daily volume of likes; (iii) Facebook daily volume of comments.

4. *Google Trends.* The time series represent the relative volume of searches that users performed at Google, related to brands' names for the period of time from November 1, 2013, to October 31, 2015.

6.2.2 Feature Engineering

Feature engineering is a process of creating features that might contain information useful for prediction. In this experiment I created features that reflect volume and direction of sales; volume, sentiment and direction of social media; presence/absence of spikes in sales and social media (section 5.2.1 and Kolchyna et al. (2016)); presence/absence of popular holidays; seasonality, such as day of the week, month, day of the month. As there is often a delay in time between the interaction on social media networks and a purchase, I also included the lagged values of variables. A full table of different types of features used in this experiment is presented in Fig. 6.2.

Since one of the main goals of this experiment is to compare the impact of different social media factors on sales forecasts, I combined the above features into following groups:

1. Sales, seasonality and holidays features (denoted as S)

2. Sales, seasonality and holidays features + Twitter features (denoted as S_T)

3. Sales, seasonality and holidays features + Facebook features (denoted as S_F)

4. Sales, seasonality and holidays features + Google features (denoted as S_G)

5. Sales, seasonality and holidays features + Twitter + Facebook + Google features (denoted as S_AllSM)

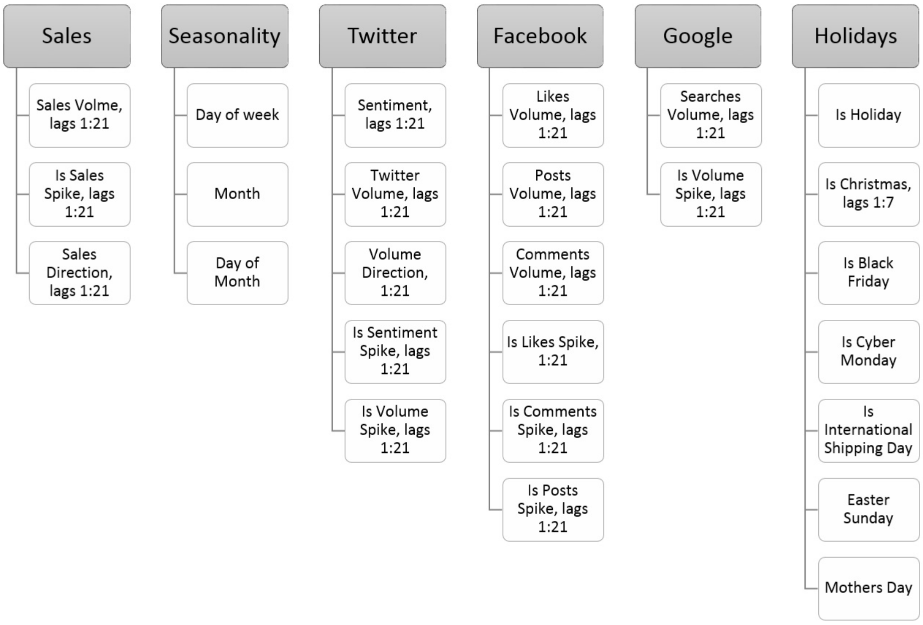

Sales
Sales Volme, lags 1:21
Is Sales Spike, lags 1:21
Sales Direction, lags 1:21
Seasonality
Day of week
Month
Day of Month
Twitter
Sentiment, lags 1:21
Twitter Volume, lags 1:21
Volume Direction, 1:21
Is Sentiment Spike, lags 1:21
Is Volume Spike, lags 1:21
Facebook
Likes Volume, lags 1:21
Posts Volume, lags 1:21
Comments Volume, lags 1:21
Is Likes Spike, 1:21
Is Comments Spike, lags 1:21
Is Posts Spike, lags 1:21
Google
Searches Volume, lags 1:21
Is Volume Spike, lags 1:21
Holidays
Is Holiday
Is Christmas, lags 1:7
Is Black Friday
Is Cyber Monday
Is International Shipping Day
Easter Sunday
Mothers Day

Figure 6.2: A list of constructed features.

6. Seasonality and holidays features + Twitter + Facebook + Google features (denoted as JustSM)

For every brand a separate model for each feature group was trained. This separation into feature groups allowed to inference which type of social interaction could be most useful for forecasting sales. For example, the impact of Twitter on sales forecasts was measured by comparing the accuracy of S_T model vs. the accuracy of S model. Notice that S model served as a benchmark as it does not includes any social media factors.

6.2.3 Feature Selection

Among the features in the above sets there might be features which are only tenuously associated with the target variable, sales volume. These irrelevant or redundant features might introduce noise. In order to select the most relevant features I used Random Forest ranking method. Random forest consists of a number of decision trees. Every node in the decision tree is a condition on a single feature. The decision at each branching point in the tree seeks to make the biggest gains towards selecting a final prediction. The ranking of the features was done based on their importance, where importance indicates the mean decrease in Mean Squared Error of prediction. For each of the six groups of features described in the previous subsection I selected the top 30 features.

Random Forest feature selection process was built into a pipeline in a way that the best features would be selected for each brand, before the training of any model. This means that each brand could have potentially different features compared to other brands, although within a brand, the features used for training of each model were the same to allow proper comparison between models.

6.3 Experimental Results

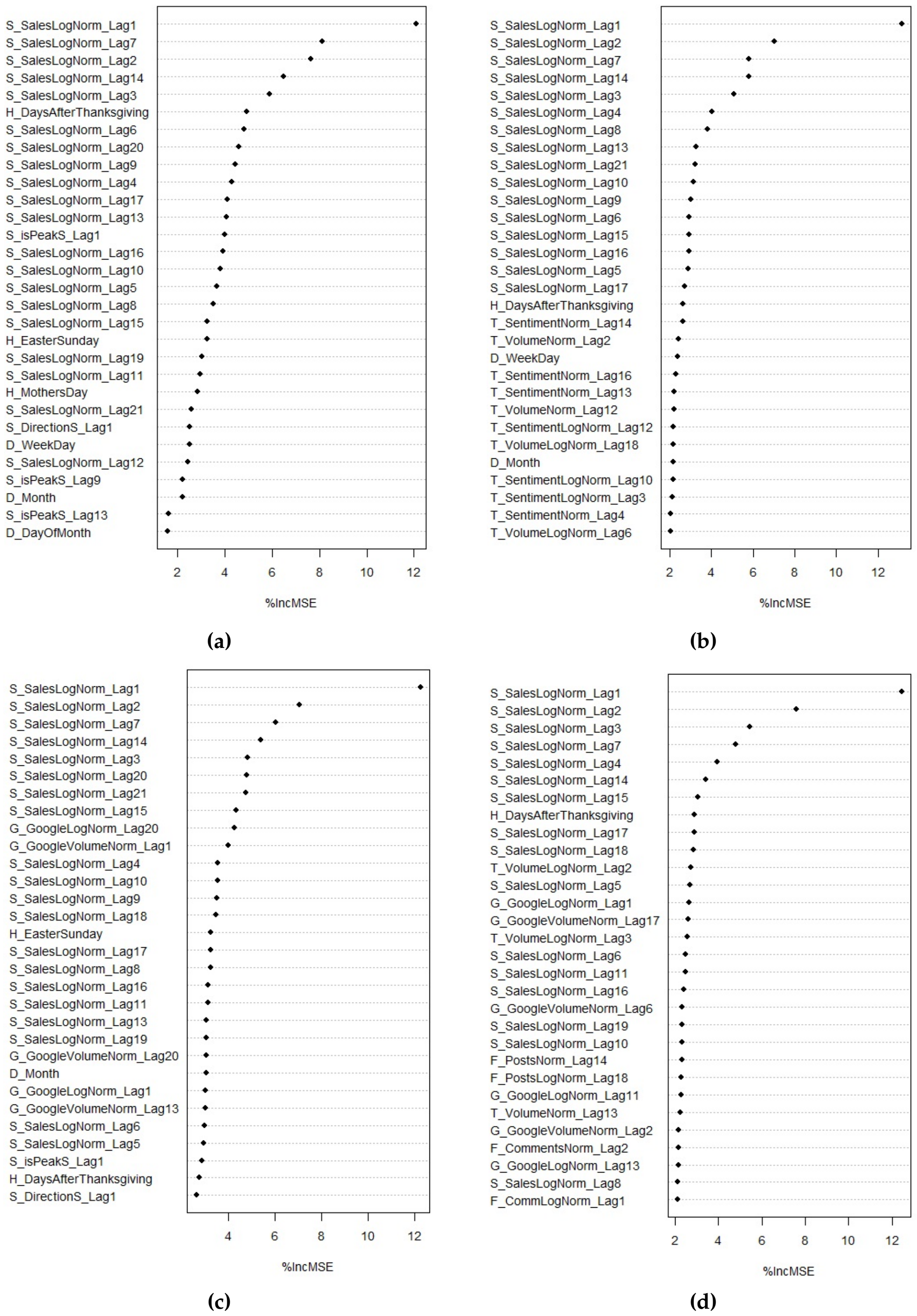
With the aim of finding more accurate prediction methods I performed analysis of 18 different retail brands. Data preparation and modelling was done individually for each brand following the methodology described in the previous section. Comparative analysis of the results across different brands allowed to answer the research questions presented at the beginning of this chapter.

6.3.1 Data preprocessing

Data exploration showed that some brands had short periods of missing data for Twitter activities. This could be related to the server being down for a short period of time. I performed data imputation for the missing periods as described in section 3.6.1. I also performed normalisation of all variables using the z-score and a log-transformation of Sales, Twitter Volume, Facebook Likes and Facebook Comments variables, as described in 3.6.2.

6.3.2 Feature Analysis

To demonstrate the results of feature selection process, in Figure 6.3 I present features chosen for brand 1 for feature groups S, S_T, S_G and S_AllSM. %IncMSE indicates the increase in Mean Squared Error as a result of a corresponding feature on the y-axis being permuted. The higher is the number, the more important is the feature. Feature names in Figure 6.3 have the following structure: *FeatureType_FeatureAndPreprocessingType_LagValue*. *FeatureType* is one of the 6 types of features presented in Fig. 6.2 (S_ stands for sales, D_ for seasonality, T_ for Twitter, F_ for Facebook, G_ for Google, H_ for holidays). *FeatureAndPreprocessingType* represents the meaning of the feature and the type of pre-processing: logging, normalisation, both or none. *LagValue* denotes the lag of the variable, which can range between 1 and 21 days.



(a)

(b)

(c)

(d)

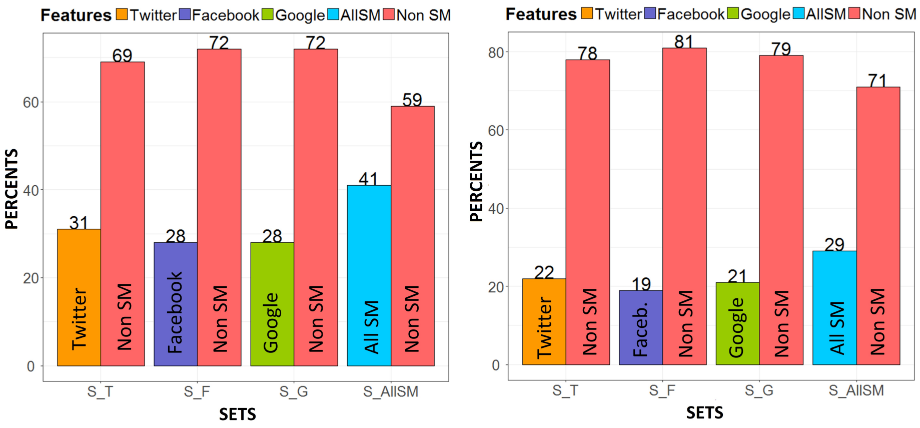
Figure 6.3: Feature importance for brand 1 for top thirty features of the feature groups: a) S, b) S_T, c) S_G, d) S_AllSM. S_ stands for sales, D_ for seasonality, T_ for Twitter, F_ for Facebook, G_ for Google, H_ for holidays.

Fig. 6.3.a displays 30 features that were selected for a group of features S (only sales related features and no social media). We see that the value of sales on the previous day S_SalesLogNorm_Lag1 is the most important feature for predicting next day sales. We also observe that the indication of the spike in sales on the previous day S_isPeakS_Lag1 and the direction of yesterday's sales S_DirectionS_Lag1 both have high importance for forecasting sales. The second most important feature is sales on the same day of the week one week before, S_SalesLogNorm_Lag7. This indicates that sales have weekly patterns, with some days of the week being more busy in terms of purchases than others. This hypothesis is also supported by the highly ranked feature D_WeekDay. Selection of the feature D_Month as well as some holiday related features indicates that sales data is seasonal: indeed, retail sales grow from September through December with the peak around Christmas, and then decline in January and February.

In Fig. 6.3.b we can see features selected for a group S_T (sales and Twitter features). While the top features are lagged sales features, we can observe that multiple Twitter features were selected (11 out of 30). This indicates that Twitter contains information about future sales. Of all Twitter features selected for brand 1, seven were Twitter sentiment features and 4 were Twitter volume features. This suggests that sentiment might be more informative than Twitter volume. When looking at the feature group S_G (sales and Google features) we see that only 5 out of 30 features were Google features, however, some of them had very high importance appearing among the top 10 features and comparable to lagged sales (Fig.6.3.d).

Results of feature selection after combining all possible features (sales, Twitter, Facebook and Google features) are presented in Fig. 6.3.e. It is important to notice that all types of features were represented in the top 30 selection (overall, 12 social media variables were selected). This indicates that Twitter, Facebook and Google data, all have the ability to explain future sales.

To present the results of different brands in a compact form I calculated the proportions of social media features selected for each feature group for each brand. The results are presented in Appendix C Figs. 8.5-8.8. Figs. 8.9-8.12 in Appendix C present the contributions that features of each type deliver to the reduction in MSE



(a)

(b)

Figure 6.4: a) Proportions of selected feature types for different feature groups averaged across brands, b) Contribution that features of each type deliver to the reduction in MSE on average across brands.

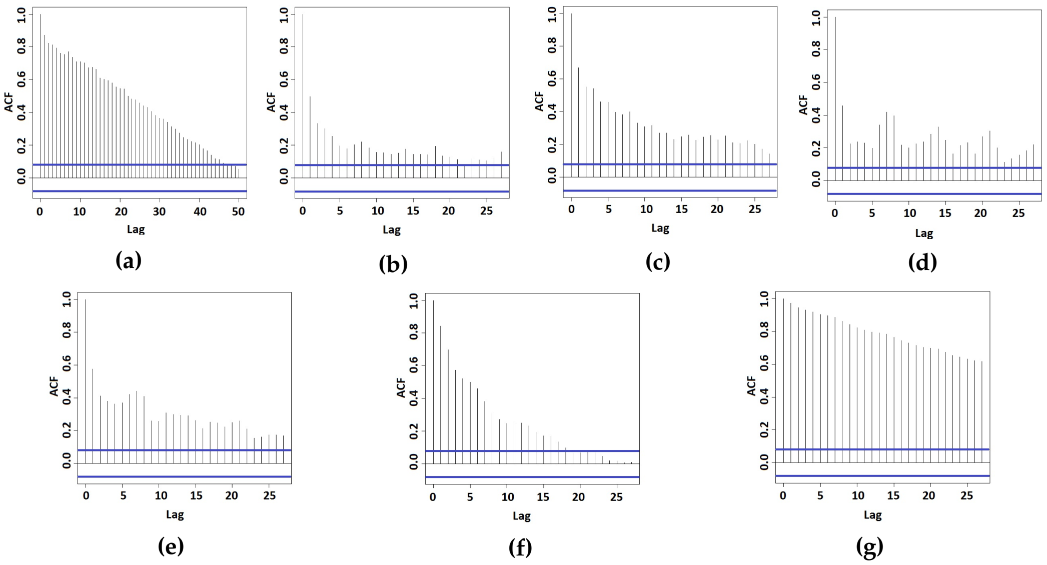
for each brand.

To summarise the results across brands I calculated proportions of different feature types for different feature groups (Fig. 6.4.a). For the feature group S_F on average 28% of features were Facebook features, for the feature group S_G on average 28% of features were Google related, for the feature group S_T on average 31% of features were Twitter related and for the feature set S_AllSM on average 41% of all features were social media features.

Fig. 6.4.b presents the average contribution across brands that features of each type deliver to the reduction in MSE. For example, for a group S_T, 31% of Twitter features (6.4.a) contributed to 22% of the reduction in MSE (6.4.b), while 69% of sales features corresponded to 78% of the reduction in MSE. These numbers mean that a Twitter feature on average contains less information about future sales than a sales feature. This outcome is not surprising since past sales are expected to be the best predictors of future sales.

6.3.3 Stationarity Analysis

In this section I performed stationarity testing for each of the time series: logged sales volume (S_VolLog), Twitter volume (T_Vol), Twitter sentiment (T_Sent), Facebook comments volume (F_Comm), logged Facebook posts volume (F_PostsLog), logged Facebook Likes volume (F_LikesLog) and logged Google searches volume (G_VolLog) as described in section 2.2.3. I first



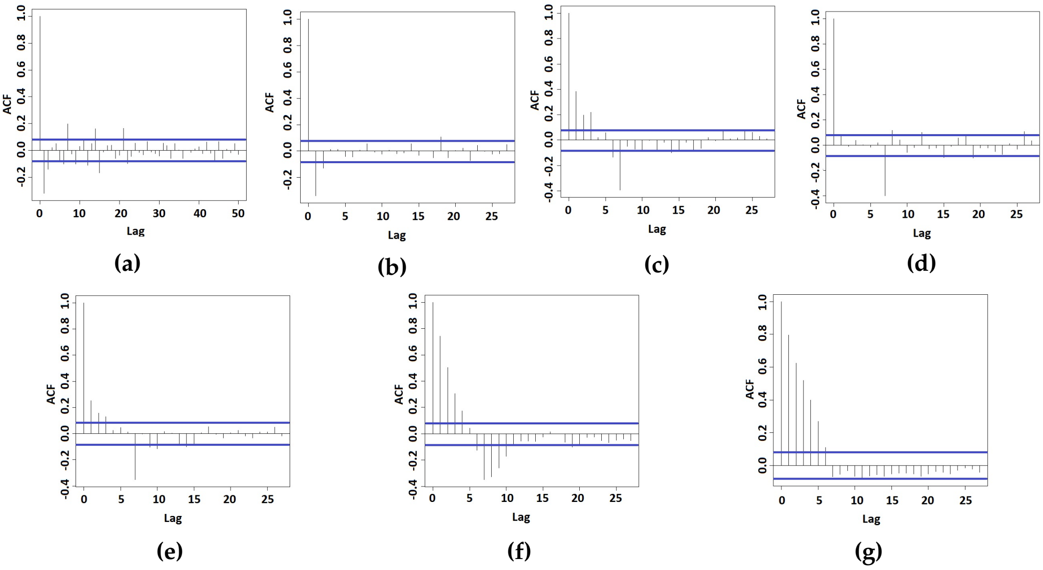
(a) (b) (c) (d) (e) (f) (g)

Figure 6.5: ACF correlograms for brand 9 for variables: a) S_VolLog, b) T_Vol, c) T_Sent, d) F_LikesLog, e) F_PostsLog, f) F_Comm, g) G_VolLog. Since the ACFs did not drop from one to zero relatively quickly for all variables, we conclude that time series are not stationary at levels.

assessed autocorrelation functions (ACF) for each variable. I also applied Kwiatkowski-Phillips-Schmidt-Shin Test (Kwiatkowski et al., 1991) in combination with Augmented Dickey-Fuller Test (Dickey and Fuller, 1979) and Phillips-Perron Test Phillips and Perron (1988). The null hypothesis of the KPSS test is that time series is stationarity, while ADF and PP test the null hypothesis of the presence of a unit root. The presence of a unit root indicates that the process is non mean-reverting (non-stationary). If the results from at least two of four tests suggested that time series was non-stationary, I accepted the non-stationarity of that variable. To demonstrate the approach I present the results of four stationarity tests for one of the brands.

Fig. 6.5 presents the correlograms of the ACF test for brand 9. We can observe strong persistent correlations for all variables that do not decline after a few lags. This means that time series require differencing, and thus, are non-stationary at levels. After performing first order differencing, correlograms spikes declined in 10 days for all variables and remained within the significance range (blue lines) (Fig. 6.6). We can conclude that the variables became stationary after the first differencing, and, therefore, are integrated of order one I(1).

Table 6.1 presents the statistics for all four tests and integration orders obtained



(a) (b) (c) (d) (e) (f) (g)

Figure 6.6: ACF correlograms for brand 9 after the first order differencing of variables: a) S_VolLog, b) T_Vol, c) T_Sent, d) F_LikesLog, e) F_PostsLog, f) F_Comm, g) G_VolLog. The ACFs declined within 10 days for all variables and remained within the significance range (blue lines). We conclude that time series became stationary after first-order differencing, therefore they are I(1).

Table 6.1: Results of stationarity tests for brand 9.

	ADF			KPSS		PP		ACF	Final
	test stats	critical values	order	p-value	order	p-value	order	order	order
S_SalesLog	-2.62	-3.41	1	<0.01	1	0.19	1	1	1
T_Volume	-4.93	-3.41	0	<0.01	1	<0.01	0	1	1
T_Sentiment	-5.17	-3.41	0	<0.01	1	0.12	1	1	1
F_Comments	-4.61	-3.41	0	<0.01	1	0.68	1	1	1
F_PostsLog	-3.93	-3.41	0	<0.01	1	<0.01	0	1	1
F_LikesLog	-3.19	-3.41	1	<0.01	1	<0.01	0	1	1
G_GoogleLog	-2.45	-3.41	1	<0.01	1	0.53	1	1	1

from each test. In ADF test, the test statistic is greater than the respective critical value for variables S_SalesLog, F_LikesLog and G_GoogleLog, thus, we cannot reject the null hypothesis of presence of a unit root at 95% confidence level for these variables. The results of KPSS test for all variables show that p-value is smaller than 0.01, meaning we should reject the null hypothesis that variables are level stationary. The results of PP test suggest that S_SalesLog, T_Sentiment, F_Comments and G_GoogleLog have unit roots. The Final order column in Table 6.1 represents the decision on the integration order of each variable based on the

Table 6.2: Integration orders for different features of each brand.

	S_Vol Log	T_Vol	T_Sent	F_Comm	F_Posts Log	F_Likes Log	G_Vol Log
Brand 1	0	1	1	0	0	0	1
Brand 10	1	1	1	1	1	1	1
Brand 12	1	1	1	1	1	1	1
Brand 14	0	0	1	0	1	1	1
Brand 15	1	1	1	1	1	1	1
Brand 17	1	1	1	0	1	1	1
Brand 18	1	0	1	1	1	0	1
Brand 2	1	1	1	0	1	1	1
Brand 20	1	1	1	1	1	1	1
Brand 21	1	1	1	1	1	1	1
Brand 22	1	1	1	0	1	1	1
Brand 24	1	1	1	1	1	1	1
Brand 3	0	0	1	0	1	1	1
Brand 4	1	0	1	1	1	1	1
Brand 6	1	0	1	0	1	1	1
Brand 7	1	1	1	0	1	1	1
Brand 8	1	0	1	0	1	1	1
Brand 9	1	1	1	1	1	1	1

outcomes of four tests.

Following the described approach I found integration orders for different variables of each brand (Table 6.2). We can see that most of the variables are integrated of order one I(1).

6.3.4 Cointegration Analysis

According to the approach discussed in section 6.2, if for any brand all variables are integrated of order one and are cointegrated I can use a Vector Error Correction Model on levels of data. To check for cointegration I run Johansen test using trace statistics as described in Section 2.2.6. The features were grouped by the type of social media so I could compare their ability to improve sales forecasts.

As an example, results of Johansen test for brand 9 and feature groups S_T and S_G are presented in Table 6.3. Testing using Johansen procedure proceeds sequentially for $r = 1,2$, etc., and the first non-rejection of the null is taken as an estimate of r. Let's look at the feature group S_T. The first hypothesis, $r = 0$, tests for the presence of cointegration between sales and Twitter variables. Since the test statistic exceeds the 1% level significantly ($75.89 > 41.07$) there is a strong evidence

Table 6.3: Results of Johansen test for brand 9: test statistics and critical values for feature groups S_T and S_G.

	S_T				S_G			
	Test	10pct	5pct	1pct	Test	10pct	5pct	1pct
r <= 2	4.53	7.52	9.24	12.97				
r <= 1	31.54	17.85	19.96	24.60	4.96	7.52	9.24	12.97
r = 0	75.89	32.00	34.91	41.07	11.94	17.85	19.96	24.60

Table 6.4: Optimal lags for different feature sets for each brand

	S_AllSM			S_T			S_F			S_G		
	Lag	Coint Order	Vars N	Lag	Coint Order	Vars N	Lag	Coint Order	Vars N	Lag	Coint Order	Vars N
Brand 10	7	6	7	7	2	3	7	3	4	9	1	2
Brand 12	3	6	7	3	2	3	7	3	4	8	1	2
Brand 15	8	5	7	14	2	3	8	2	4	15	0	2
Brand 17	7	3	6	4	2	3	3	3	3	15	1	2
Brand 18	9	4	5	7	1	2	13	2	3	15	1	2
Brand 2	8	5	6	7	2	3	14	2	3	15	0	2
Brand 20	7	5	7	4	2	3	7	3	4	16	1	2
Brand 21	2	6	7	8	2	3	2	3	4	15	1	2
Brand 22	8	4	6	7	2	3	16	1	3	15	1	2
Brand 24	8	3	7	4	2	3	2	3	4	15	1	2
Brand 4	9	3	6	7	1	2	14	3	4	16	1	2
Brand 6	8	3	5	7	1	2	7	2	3	15	1	2
Brand 7	7	3	6	7	1	3	8	2	3	15	1	2
Brand 8	7	2	5	14	1	2	15	1	3	16	1	2
Brand 9	3	5	7	7	2	3	9	3	4	16	0	2

to reject the null hypothesis of no cointegration. The second test for $r <= 1$ against the alternative hypothesis of $r > 1$ also provides clear evidence to reject $r <= 1$ since the test statistic exceeds the 1% level significantly (31.54 > 24.60). The final test for $r <= 2$ against $r > 2$ does not allow to reject the null hypothesis that $r <= 2$ since test statistic is smaller than critical value even at 10% level of significance (4.53 < 7.52). I conclude that the cointegration rank between sales and Twitter variables is 2. Following the same procedure I conclude that there is no cointegration between sales and Google variable (11.94 < 17.85 < 19.96 < 24.60).

Table 6.4 presents the number of lags, cointegration order and number of variables for all brands. The most appropriate number of lags was found by minimizing Akaike's information criterion (AIC). The lag k, i.e., means that lagged variables up to k should be included in the cointegration relationship. If the order of integration is higher than zero, all of those lagged variables should be used in estimating the VECM. Sales data for brands 1, 14 and 3 is stationary at levels, therefore, they were excluded from cointegration analysis. For the mentioned brands only MLR and ANN were trained. For other brands, in order to satisfy the requirement of having all variables of the same order, I excluded those variables that have the integration order of zero. For example, the variable F_Comm for brand 17 is I(0), and, therefore, it was excluded from the cointegration analysis and from the VECM modelling.

According to Table 6.4 the cointegration ranks for all brands are higher than zero and less than the number of variables. This indicates that the series are cointegrated among the variables for all brands and a VECM can be used to estimate the long-run relationship between variables.

6.3.5 Modelling

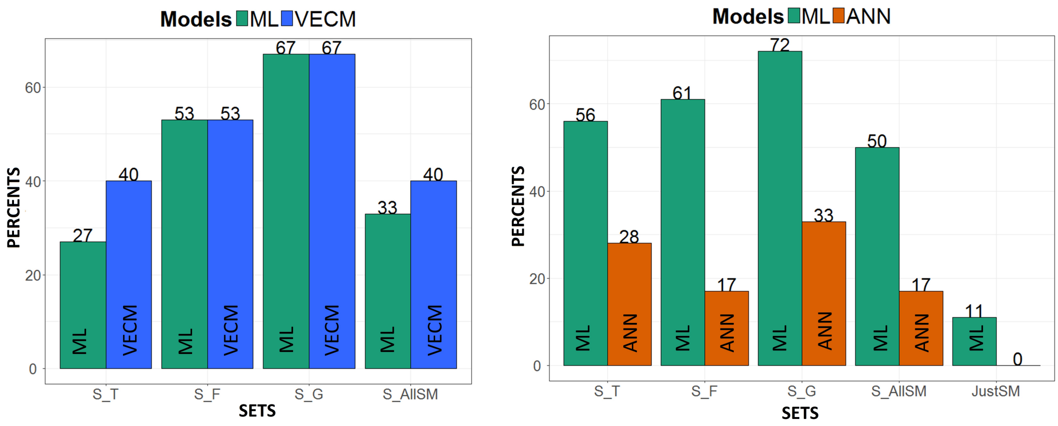
To answer the questions stated at the beginning of the chapter I performed the following steps:

1. Used the autoregressive model trained on lagged sales data only as a benchmark for each brand (see example set of features in Fig. 6.3.a).

2. Compared MLR vs. VECM using the Feature Set 1 for brands that have social media variables cointegrated with sales. The Feature Set 1 consists of features that are all lagged of the same order, selected through the lag selection process (see Table 6.4 for optimal lags oders). For example, if selected lag equals five for the feature group S_T, it means that sales, Twitter volume and Twitter sentiment variables all were lagged 5 times and included in the model.

3. Compared MLR vs. ANN using the Feature Set 2 for all brands. The Feature Set 2 consists of the variables selected by Random Forest as described in section 6.3.2.

4. Selected the best model for each brand from the models generated during steps 1 - 3. The best model is the one that has the biggest relative MAPE reduction compared to the benchmark. Reduction in MAPE is calculated as



(a)

(b)

Figure 6.7: Percentage of brands for which performance improved when using social media features compared to the benchmark based on MAPE for a) Feature Set 1, b) Feature Set 2. Different feature groups are presented along X-axis. Bar colors represent different model types.

follows:

$$\text{MAPEReduction} = \frac{(\text{MAPEBenchmark} - \text{MAPEAlternative})}{\text{MAPEBenchmark}} * 100 \qquad (6.1)$$

Since predicting the actual sales figure is a very complex task I also predicted the target times series' immediate direction, that is, whether the next day sales value is higher or lower than the last observation. For this binary classification task I compared the accuracy of classification of logistic regression and ANN using the Feature Set 2.

In the current experiment I used the standard forward ANN with three layers, one input layer, one hidden layer and one output layer. The thirty features selected through Random Forest ranking process served as the input nodes. The hidden layer was comprised of twenty nodes and the output layer had one neuron, the value of sales time series at $t + 1$ time moment. I used a linear activation function for the input and output layers and a sigmoid function for the hidden layer.

6.3.6 Forecasts of Sales Volume

MLR vs. VECM on Feature Set 1. Table 6.5 presents the results of MAPE calculations for brands that have cointegrating relationships between the variables. The first column of the table is the MAPE for the benchmark. The table allows to

Table 6.5: MAPE results of predicting sales volume using MLR and VECM on the feature set 1. The first line for each brand presents MAPE results for different features groups (%), the second line presents relative percentage improvement over the benchmark. Light and dark colors represent improvements by more than 1% and 3%, respectively.

		Multiple Regression				Error Correction Model				Best
	S	S_T	S_F	S_G	S_All	S_T	S_F	S_G	S_All	Model
Brand 10	18.3	18.1	17.5	17.8	18.0	18.2	17.4	17.9	18.2	VECM_F
		(1.2)	(4.6)	(2.7)	(1.7)	(0.8)	(4.9)	(2.4)	(0.5)	
Brand 12	51.0	50.5	47.2	47.6	51.4	52.6	44.9	47.4	48.2	VECM_F
		(1.0)	(7.5)	(6.8)	(-0.7)	(-3.0)	(12.0)	(7.2)	(5.5)	
Brand 15	21.2	25.0	25.4	23.5	24.3	25.3	26.9	-	25.4	MLR_S
		(-17.9)	(-19.8)	(-10.9)	(-14.6)	(-19.3)	(-26.9)	-	(-19.8)	
Brand 17	22.9	26.5	22.1	23.0	20.6	22.2	21.4	21.7	20.6	VECM_All
		(-15.7)	(3.8)	(-0.5)	(10.0)	(3.2)	(6.6)	(5.1)	(10.3)	
Brand 18	25.3	25.9	29.3	25.6	29.1	23.0	28.5	23.9	22.5	VECM_All
		(-2.3)	(-15.8)	(-0.9)	(-14.9)	(9.4)	(-12.3)	(5.8)	(11.2)	
Brand 2	16.3	16.7	15.8	15.6	15.8	17.2	16.4	-	16.1	MLR_G
		(-2.4)	(2.8)	(4.1)	(2.9)	(-5.6)	(-0.6)	-	(1.0)	
Brand 20	26.3	26.1	27.6	25.6	27.8	25.2	24.4	22.9	27.7	VECM_G
		(0.8)	(-4.9)	(2.8)	(-5.6)	(4.1)	(7.4)	(13.2)	(-5.3)	
Brand 21	151.5	158.4	155.4	154.4	149.9	153.4	148.9	155.5	142.3	VECM_All
		(-4.5)	(-2.6)	(-1.9)	(1.1)	(-1.2)	(1.7)	(-2.6)	(6.1)	
Brand 22	25.7	27.0	24.2	25.0	27.4	27.0	23.5	24.3	26.9	VECM_F
		(-5.2)	(5.6)	(2.7)	(-6.8)	(-5.0)	(8.3)	(5.3)	(-5.0)	
Brand 24	24.3	25.0	24.7	25.3	25.5	25.1	26.9	25.0	25.7	MLR_S
		(-3.0)	(-1.6)	(-4.1)	(-4.8)	(-3.1)	(-10.7)	(-2.7)	(-5.6)	
Brand 4	14.4	15.2	15.5	14.3	14.5	14.2	15.1	14.0	14.3	VECM_G
		(-5.4)	(-7.4)	(1.1)	(-0.3)	(1.4)	(-4.8)	(3.1)	(1.0)	
Brand 6	40.6	38.0	37.9	40.1	40.9	36.8	38.2	37.6	36.1	VECM_All
		(6.3)	(6.5)	(1.1)	(-0.9)	(9.4)	(5.8)	(7.4)	(11.0)	
Brand 7	7.8	9.6	8.6	7.6	9.6	9.7	8.6	7.5	9.5	VECM_G
		(-22.9)	(-10.0)	(2.3)	(-22.3)	(-23.3)	(-9.7)	(3.9)	(-21.9)	
Brand 8	520.0	547.7	507.6	476.8	421.4	530.4	492.1	467.6	429.3	MLR_All
		(-5.3)	(2.4)	(8.3)	(19.0)	(-2.0)	(5.4)	(10.1)	(17.4)	
Brand 9	14.4	12.9	14.1	13.6	14.6	13.5	14.0	-	15.0	VECM_T
		(10.5)	(1.7)	(5.4)	(-1.2)	(6.2)	(2.8)	-	(-3.9)	
Average	65.3	68.2	64.9	**62.4**	**59.4**	66.2	**63.1**	**61.3**	**58.5**	
MAPE		(-4.3)	(0.76)	**(4.6)**	**(9.2)**	(-1.2)	**(3.5)**	**(6.3)**	**(10.6)**	

compare how adding additional variables changes the performance of the models as well as comparing MLR model vs. the VECM for different feature groups. The first row of values for each brand indicates the actual MAPE, while the numbers in the brackets in the second row present a relative percentage improvement over the benchmark using Eq. 6.1 (larger values indicate greater improvement). The light grey color highlights models that have relative percentage improvement of MAPE of at least 1% and the dark grey color highlights the models that improved the MAPE by more than 3% compared to the benchmark. The last column in the table shows the model and feature group that performed best in terms of MAPE.

Table 6.6: MdAPE results of predicting sales volume using MLR and VECM on the feature set 1. The first line for each brand presents MdAPE results for different features groups (%), the second line presents relative percentage improvement over the benchmark. Light and dark colors represent improvements by more than 1% and 3%, respectively.

		Multiple Regression				Error Correction Model				Best
	S	S_T	S_F	S_G	S_All	S_T	S_F	S_G	S_All	Model
Brand 10	15.5	15 (3.4)	14.7 (5.1)	14.2 (8.7)	14.2 (8.6)	14.8 (4.8)	15.0 (3.7)	14.8 (4.5)	15.7 (-1.4)	MLR_G
Brand 12	37.2	34.3 (7.8)	31.5 (15.4)	33.3 (10.6)	34.6 (7.1)	34.2 (8.2)	32.4 (13.1)	31.4 (15.7)	32.1 (13.8)	VECM_G
Brand 15	15.2	17.2 (-13.0)	19.4 (-28.0)	16.3 (-7.2)	17.4 (-14.5)	17.6 (-15.8)	20.7 (-36.1)	- -	18.2 (-19.8)	MLR_S
Brand 17	18.9	18.1 (4.2)	16.1 (14.7)	18.0 (4.8)	16.3 (14.0)	16.8 (11.2)	16.4 (13.3)	18.7 (1.2)	16.0 (15.7)	VECM_All
Brand 18	20.7	18.9 (8.5)	21.7 (-4.8)	20.7 (0.0)	21.2 (-2.7)	18.3 (11.4)	22.0 (-6.2)	19.9 (3.9)	18.2 (11.8)	VECM_All
Brand 2	11.7	10.9 (6.5)	10.3 (11.7)	10.5 (9.7)	10.4 (11.2)	10.5 (9.9)	10.7 (8.1)	- -	10.2 (12.4)	VECM_All
Brand 20	18.9	20.3 (-7.6)	21.3 (-12.9)	20.5 (-8.7)	19.6 (-3.6)	19.6 (-3.7)	19.‘ (-1.2)	17.2 (8.6)	20.0 (-5.9)	VECM_G
Brand 21	12.7	12.8 (-1.2)	11.9 (6.4)	12.0 (5.5)	13.3 (-4.8)	12.5 (1.5)	12.2 (3.9)	12.3 (3.2)	14.8 (-16.7)	MLR_F
Brand 22	19.4	21.2 (-9.3)	18.3 (5.7)	20.2 (-4.1)	18.8 (3.2)	21.1 (-8.9)	17.3 (10.8)	19.9 (-3.0)	18.2 (5.8)	VECM_F
Brand 24	18.0	15.9 (11.4)	17.1 (5.0)	17.5 (2.5)	16.5 (8.0)	15.9 (11.6)	18.8 (-4.7)	16.6 (7.6)	17.5 (2.9)	VECM_T
Brand 4	13.5	13.1 (3.2)	12.2 (9.5)	10.0 (25.9)	11.0 (18.4)	11.4 (15.6)	10.6 (21.5)	10.3 (23.9)	11.0 (18.3)	MLR_G
Brand 6	17.7	13.7 (22.6)	14.0 (21.2)	16.7 (5.7)	16.2 (8.8)	14.5 (18.2)	14.8 (16.6)	16.0 (9.7)	13.9 (21.7)	MLR_T
Brand 7	6.2	7.3 (-16.5)	6.4 (-2.0)	5.9 (6.1)	6.9 (-9.9)	7.6 (-21.2)	6.4 (-1.9)	5.6 (10.6)	7.3 (-17.3)	VECM_G
Brand 8	36.0	38.7 (-7.4)	37.7 (-4.7)	36.0 (0.2)	41.1 (-14.0)	36.0 (0.1)	38.1 (-5.8)	36.6 (-1.5)	41.4 (-15.0)	MLR_S
Brand 9	10.1	9.4 (7.1)	10.9 (-8.1)	9.2 (8.5)	11.1 (-9.5)	9.3 (8.3)	11.0 (-8.6)	- -	11.7 (-15.8)	MLR_G
Median MdAPE	17.7	**15.9** **(10.1)**	**16.1** **(9.0)**	**16.7** **(5.7)**	**16.3** **(8.2)**	**15.9** **(10.3)**	**16.4** **(7.5)**	**16.1** **(9.0)**	**16.0** **(10.0)**	

From Table 6.5 we can see that the VECM part of the Table has more cell coloured in grey color than the MLR part of the table. This suggests that the improvement over the benchmark occurs more often for the VECM rather then for MLR. Also, the improvements achieved by VECM are higher (over 3%) for the majority of the brands. For example, for brand 17 we can see that S_F and S_All feature groups performed better than the benchmark for the MLR model by 3.8% and 10.0%, respectively. When the VECM was used, the improvement can be observed for all four feature groups, including Twitter and Google. Moreover, the improvement for Facebook group and combined social media features group is higher than that for MLR model, 6.6% and 10.3%, respectively. Similar conclusion can be derived when looking at the average MAPE (the last row in the table). We can see that on average across the brands, there is an improvement in MAPE for S_F, S_G and S_All groups and the improvement is higher for the VECM model. We do not observe the improvement for the S_T group when the results are averaged across different brands. Overall, for 8 out of 15 brands the improvement by more than 3% was achieved by adding social media features when using MLR, and for 12 out of 15 brands the MAPE was reduced when using the VECM with social media variables.

For brands number 21 and 8 MAPE values are very high. This is related to the occurrence of large spikes in sales data. In this case Median Absolute Percentage Error (MdAPE) is a better metric than MAPE. The results of MdAPE are presented in Table 6.6. In terms of MdAPE the difference in performance between the benchmark and models enhanced by social media features is even greater, for both, MLR model and VECM. The median MdAPE across brands (the last row of Table 6.6) shows that all social media feature groups (including S_T) improve the MdAPE, with the improvements ranging from 5.7% to 10.3%. The median MdAPE for VECM showed higher improvements over the benchmark compared to MLR.

The performance of both models and different feature groups across brands is summarised in the plot of histograms in Fig. 6.7.a. The bars represent the percentage of brands for which the MAPE was reduced by adding social media features when compared to the benchmark. Green color indicates the MLR model and the blue color represents the VECM. The performance is also separated for different feature groups. For example, the green bar in Fig. 6.7.a in a Twitter group with the number 27 indicates that the MLR model with Twitter features performed better than the benchmark for 27% of brands. We can see that the bars for VECM are higher than those for MLR for S_T and AllSM feature groups. This means that the VECM lead to improved results more often than MLR for those feature sets. Also, we can see that bars in the S_G group are higher than the bars in other groups for both models, indicating that adding Google features helps to model sales more accurately than when using other social media features.

Based on the Feature Set 1, we can conclude that including Twitter, Facebook

Table 6.7: MAPE results of predicting sales volume using MLR and ANN on the feature set 2. The first line for each brand presents MAPE results for different features groups (%), the second line presents relative percentage improvement over the benchmark. Light and dark colors represent improvements by more than 1% and 3%, respectively.

	Multiple Regression						Artificial Neural Networks						Best
	S	S_T	S_F	S_G	S_All	JustSM	S	S_T	S_F	S_G	S_All	JustSM	Model
Brand 1	16.9	16.8	15.1	15.8)	16.5	34.6	16.7	18.0	16.5	16.6	17.7	46.5	MLR_F
		(0.7)	(10.9)	(6.3)	(2.3)	(-104.5)		(-7.8)	(1.3)	(0.8)	(-6.2)	(-178.3)	
Brand 10	18.2	18.7	17.5	16.7	17.3	25.9	17.5	23.5	20.9	21.6	24.8	44.5	MLR_G
		(-2.4)	(4.0)	(8.2)	(4.9)	(-41.9)		(-34.0)	(-19.1)	(-23.4)	(-41.8)	(-154.5)	
Brand 12	51.0	46.3	48.2	48.3	48.9	82.8	48.4	45.4	50.7	51.4	93.8	223.1	ANN_T
		(9.3)	(5.4)	(5.2)	(4.1)	(-62.4)		(6.3)	(-4.6)	(-6.2)	(-93.7)	(-360.8)	
Brand 14	12.6	12.4	12.3	11.2	12.8	13.6	12.8	13.5	12.5	12.1	11.4	79	MLR_G
		(1.6)	(2.4)	(11.1)	(-1.6)	(-7.9)		(-5.5)	(2.3)	(4.8)	(10.7)	(-519.3)	
Brand 15	24.3	21.8	21.8	21.8	21.6	41.0	24.3	28.2	42.5	23.1	43.2	104.9	MLR_All
		(10.2)	(10.3)	(10.4)	(11.0)	(-68.9)		(-15.8)	(-74.9)	(5.0)	(-77.9)	(-331.5)	
Brand 17	22.7	26.8	22.9	22.3	25.1	24.1	23.3	47.9	36.0	22.5	40.4	194.4	MLR_G
		(-17.8)	(-0.9)	(2.0)	(-10.6)	(-6.2)		(-105.6)	(-54.5)	(3.5)	(-73.1)	(-734.0)	
Brand 18	24.7	23.5	23.8	23.9	25.9	39.6	25.5	34.3	32.6	28.2	27.5	95.7	MLR_T
		(4.8)	(3.9)	(3.3)	(-4.7)	(-60.3)		(-34.6)	(-27.9)	(-10.7)	(-7.7)	(-275.2)	
Brand 2	15.8	14.9	15.0	15.3	15.6	23.2	16.4	18.2	19.6	22.3	26.3	63.7	MLR_T
		(5.5)	(5.2)	(3.1)	(0.8)	(-46.8)		(-11.3)	(-19.9)	(-36.4)	(-60.5)	(-289.4)	
Brand 20	23.4	25.8	27.4	23.7	25.3	25.9	26.2	25.4	49.6	26.2	28.9	87.8	MLR_S
		(-10.4)	(-17.0)	(-1.3)	(-8.3)	(-10.7)		(3.1)	(-89.4)	(-0.1)	(-10.4)	(-235.2)	
Brand 21	136.1	147.1	142.9	136.0	160.0	187.4	100.9	135.4	151.3	300.4	236.3	221.1	ANN_S
		(-8.1)	(-5.0)	(0.1)	(-17.3)	(-37.7)		(-34.2)	(-49.9)	(-197.8)	(-134.2)	(-119.2)	
Brand 22	24.2	24.6	21.9	22.7	23.1	23.1	23.8	25.9	23.8	23.4	24.6	33.4	MLR_F
		(-1.6)	(9.53)	(6.08)	(4.66)	(4.63)		(-9.06)	(-0.28)	(1.62)	(-3.41)	(-40.34)	
Brand 24	25.2	26.2	24.2	25.0	26.5	29.2	30.1	27.3	26.6	36.7	25.9	329.9	MLR_F
		(-3.8)	(3.9)	(0.6)	(-5.4)	(-15.9)		(9.4)	(11.6)	(-22.0)	(13.8)	(-996.3)	
Brand 3	18.5	19.5	32.0	18.9	162.2	21.5	18.7	19.9	31.6	18.9	36.2	44.8	ANN_S
		(-5.4)	(-73.0)	(-2.2)	(-776.8)	(-16.2)		(-6.4)	(-69.0)	(-1.1)	(-93.6)	(-139.6)	
Brand 4	14.0	13.3	13.7	12.8	13.8	16.0	15.4	15.3	16.3	15.4	13.9	143.5	MLR_G
		(5.0)	(2.1)	(8.6)	(1.4)	(-14.3)		(0.7)	(-5.5)	(0.2)	(9.7)	(-831.7)	
Brand 6	43.9	40.1	40.8	39.8	39.5	82.5	50.9	49.1	54.0	46.1	57.1	130.9	MLR_All
		(8.6)	(7.0)	(9.3)	(10.0)	(-88.3)		(3.6)	(-5.9)	(9.5)	(-12.0)	(-157.0)	
Brand 7	7.5	7.4	7.6	6.8	7.0	11.6	6.7	7.6	9.1	7.2	7.5	24.4	MLR_G
		(1.3)	(-1.3)	(9.3)	(6.7)	(-54.7)		(-13.4)	(-35.8)	(-7.5)	(-11.9)	(-264.2)	
Brand 8	474.0	456.5	493.8	524.7	508.7	332.5	378.1	352.4	644.3	284.3	366.2	487.1	ANN_G
		(3.7)	(-4.2)	(-10.7)	(-7.3)	(29.9)		(6.8)	(-70.4)	(24.8)	(3.2)	(-28.8)	
Brand 9	14.2	13.7	14.5	13.1	13.1	40.8	13.6	28.7	16.0	31.6	28.0	47.5	MLR_G
		(3.7)	(-1.5)	(8.3)	(7.8)	(-186.1)		(-112.1)	(-17.8)	(-133.5)	(-106.9)	(-250.3)	
Mean	53.8	**53.1**	55.3	55.5	64.6	58.6	47.2	50.9	69.7	54.9	61.7	133.5	
		(1.3)	(-2.8)	(-3.2)	(-20.1)	(-9.1)		(-7.8)	(-47.7)	(-16.4)	(-30.7)	(-182.9)	

and Google features adds useful information to the model and allows to improve sales forecasts. VECM often performs better than a simple MLR indicating that there is a long-run relationship between sales and social media variables.

MLR vs. ANN on Feature Set 2. A Feature Set 2 generated by Random Forest is used to compare the performance of the MLR against the ANN. The results of MAPE are presented in Table 6.7. As in Table 6.5, the light grey and dark grey colors indicate the relative percentage improvement of MAPE compared to benchmark by more than 1% and 3%, respectively.

We observe that all types of social media features lead to improvements of MAPE, separately or jointly, for all brands except brands number 20, 21 and 3, for which the benchmark model is the best performing model. We can see that MAPE values are in general lower for the Feature Set 2 than those for the Feature Set 1. For example, the average MAPE for the benchmark (the first column) is 53.8% for the Feature Set 2, while it is 65.6% for the Feature Set 1. We can also see that when MAPE is decreased after adding social media information, the improvement is generally higher for Feature Set 2 when compared to the Feature Set 1. This means that the Random Forest feature selection method is more effective in choosing the most informative features.

The results indicate that the MLR model performs better than the ANN for most of the brands. This is a surprising result since ANN was expected to capture non-linear relationships between sales and social media variables. The worse performance of ANN compared to MLR might be related to the over-fitting in the neural network due to the small number of training data (between 532 and 732 points).

While the feature group JustSM on average generates larger errors than models trained with other features, for many brands performance of a model trained only on social media data is comparable with performances of other models, for example, this is true for brands 10, 14, 17, 2, 20, 3, 4 and 7. Interestingly, for brand 22, the model trained only on social media data produced forecasts even better than the autoregressive model (MAPE of 23.1% vs. MAPE of 24.2%). As an example, I plotted forecasts for brands 22 and 4 in Figs. 6.8 and 6.9, respectively. From the figures we can see that models trained only on social media performed almost as good as the best performing model.

I also provide the results for MdAPE metric that is more stable to the impact of outliers in the data. The results are presented in Table 6.8 and are very similar to those of MAPE. We can see that on average, the models based on sales data enhanced by Facebook and Google, jointly or independently, achieve lower MdAPE than the models only trained on sales data (median of MdAPE across brands).

Table 6.9 shows the results for RelMAE metric. The values below 1 indicate

Table 6.8: MdAPE results of predicting sales volume using MLR and ANN on the feature set 2. The first line for each brand presents MdAPE results for different features groups (%), the second line presents relative percentage improvement over the benchmark. Light and dark colors represent improvements by more than 1% and 3%, respectively.

	Multiple Regression						Artificial Neural Networks						Best
	S	S_T	S_F	S_G	S_All	JustSM	S	S_T	S_F	S_G	S_All	JustSM	Model
Brand 1	13.1	12.56	11.03	12.27	10.81	20.38	12.59	14.21	11.51	11.29	11.77	39.87	MLR_All
		(4.13)	(15.81)	(6.34)	(17.47)	(-55.54)		(-12.9)	(8.56)	(10.33)	(6.56)	(-216.65)	
Brand 10	13.2	16.4	15.2	12.9	13.7	24.8	14.4	20.4	17.9	16.3	21.1	39.4	MLR_G
		(-24.0)	(-15.3)	(2.1)	(-4.0)	(-87.4)		(-41.4)	(-24.0)	(-13.2)	(-46.7)	(-173.7)	
Brand 12	37.8	36.8	36.8	38.2	36.0	51.3	32.5	31.2	35.3	35.8	57.3	67.4	ANN_T
		(2.6)	(2.7)	(-0.9)	(4.7)	(-35.6)		(3.8)	(-8.8)	(-10.2)	(-76.6)	(-107.7)	
Brand 14	9.2	10.5	9.9	9.5	10.1	9.4	11.0	11.3	10.7	10.7	9.4	12.0	MLR_S
		(-14.3)	(-7.3)	(-3.5)	(-9.0)	(-2.3)		(-2.8)	(2.8)	(2.3)	(14.2)	(-9.3)	
Brand 15	15.7	14.7	14.5	14.9	15.8	37.5	17.8	21.4	37.2	15.4	38.0	65.6	MLR_F
		(6.8)	(7.9)	(5.5)	(-0.4)	(-138.1)		(-20.3)	(-109.6)	(13.2)	(-113.9)	(-269.0)	
Brand 17	17.7	18.8	20.2	16.6	18.4	16.8	18.0	33.84	43.2	15.6	30.3	51.4	ANN_G
		(-6.1)	(-14.4)	(6.3)	(-3.9)	(5.2)		(-87.6)	(-139.7)	(13.7)	(-68.0)	(-184.9)	
Brand 18	20.6	21.2	21.1	17.9	23.5	33.9	22.3	25.1	24.8	22.9	19.6	54.1	MLR_G
		(-2.6)	(-2.2)	(13.1)	(-13.8)	(-64.5)		(-12.6)	(-11.1)	(-2.7)	(12.3)	(-142.6)	
Brand 2	11.0	10.4	9.3	9.6	10.2	19.1	11.7	12.8	14.0	17.0	21.3	25.8	MLR_F
		(6.3)	(15.7)	(12.7)	(7.3)	(-73.3)		(-10.2)	(-19.7)	(-45.6)	(-82.7)	(-121.3)	
Brand 20	18.7	21.8	21.3	18.1	20.4	22.7	19.2	19.8	30.4	19.9	20.5	33.7	MLR_G
		(-16.5)	(-13.5)	(3.4)	(-8.7)	(-21.0)		(-3.3)	(-58.5)	(-3.8)	(-7.0)	(-76.0)	
Brand 21	12.3	12.6	12.0	13.0	11.5	24.6	11.8	15.8	15.6	27.1	15.6	30.3	MLR_All
		(-2.7)	(2.5)	(-5.6)	(6.6)	(-100.5)		(-33.8)	(-31.7)	(-128.7)	(-31.6)	(-155.8)	
Brand 22	17.3	17.2	15.5	16.7	14.1	20.5	16.8	20.9	18.8	19.9	19.3	28.2	MLR_All
		(0.2)	(10.5)	(3.5)	(18.2)	(-18.7)		(-24.9)	(-11.9)	(-18.6)	(-14.9)	(-68.5)	
Brand 24	20.8	19.3	18.3	18.8	17.8	24.0	21	18.7	19.9	21.5	20.4	39.5	MLR_All
		(7.1)	(12.1)	(9.6)	(14.2)	(-15.8)		(11.0)	(5.2)	(-2.6)	(3.1)	(-88.0)	
Brand 3	16.5	17.1	17.1	17.2	16.7	18.9	16.1	17.8	21.8	15.5	17.9	39.2	ANN_G
		(-4.2)	(-4.0)	(-4.7)	(-1.3)	(-14.8)		(-10.2)	(-35.2)	(3.8)	(-10.6)	(-142.6)	
Brand 4	9.2	10.4	10.1	9.8	9.7	15.2	13.9	12.4	12.8	13.8	11.6	20.8	MLR_S
		(-13.0)	(-9.7)	(-6.1)	(-5.3)	(-65.4)		(11.0)	(8.0)	(0.3)	(16.7)	(-50.2)	
Brand 6	20.6	20.5	21.3	22.6	21.1	39.7	16.3	20.7	18.5	16.4	20.5	63.2	ANN_S
		(0.5)	(-3.5)	(-9.4)	(-2.2)	(-92.3)		(-26.6)	(-13.5)	(-0.6)	(-25.8)	(-287.4)	
Brand 7	5.6	5.7	6.2	5.3	5.5	10.2	5.5	6.4	7.7	5.4	5.7	12.3	MLR_G
		(-1.2)	(-10.3)	(6.1)	(2.0)	(-80.8)		(-15.7)	(-38.3)	(2.2)	(-3.6)	(-121.5)	
Brand 8	36.9	38.5	35.9	37.1	40.5	55.9	38.5	40.3	52.5	43.2	41.2	56.2	MLR_F
		(-4.4)	(2.6)	(-0.4)	(-9.7)	(-51.6)		(-4.6)	(-36.4)	(-12.2)	(-6.9)	(-46.0)	
Brand 9	11.1	11.9	11.2	9.3	9.0	40.8	10.8	22.6	12.8	28.0	24.1	47.9	MLR_All
		(-6.8)	(-0.9)	(16.6)	(18.9)	(-267.2)		(-110.0)	(-18.8)	(-160.6)	(-123.4)	(-344.3)	
Median	16.1	16.8	**15.3**	**15.7**	**15.0**	23.4	16.2	20.1	18.6	16.7	20.4	39.4	
		(-4.1)	**(6.7)**	**(2.3)**	**(7.1)**	(-45.1)		(-23.8)	(-14.8)	(-2.8)	(-25.9)	(-143.0)	

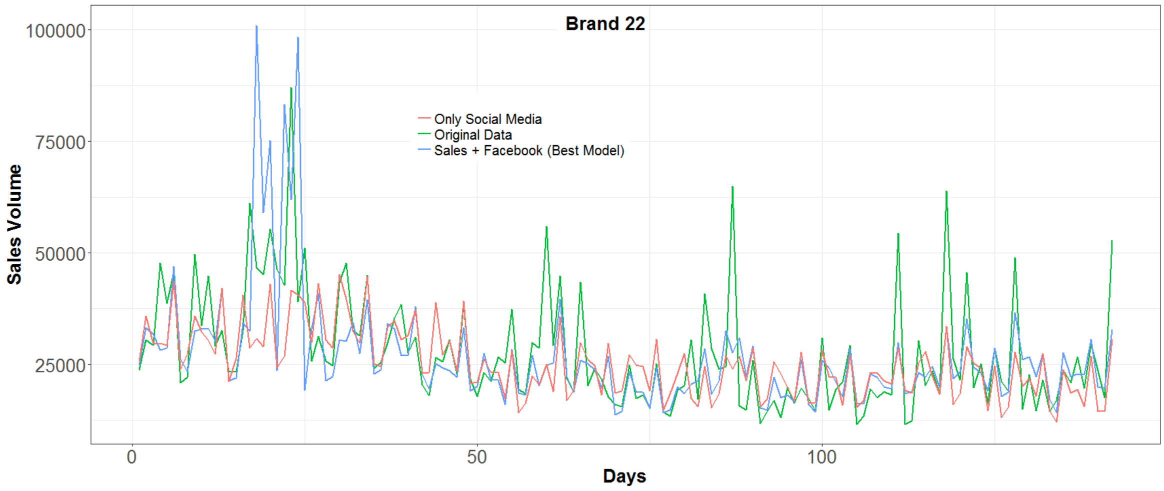


Figure 6.8: Plots of original sales data and sales forecasts for brand 22.

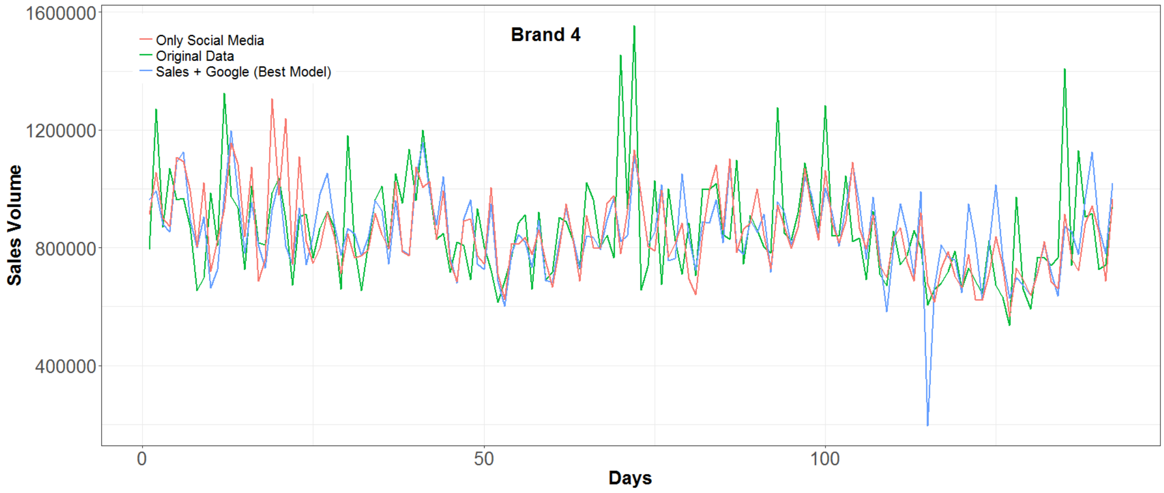


Figure 6.9: Plots of original sales data and sales forecasts for brand 4.

improved forecasts compared to the benchmark. We can see, that for the majority of the brands there is an improvement of forecasts when adding Twitter, Facebook, Google and their combination, and that the MLR model performs better than the ANN.

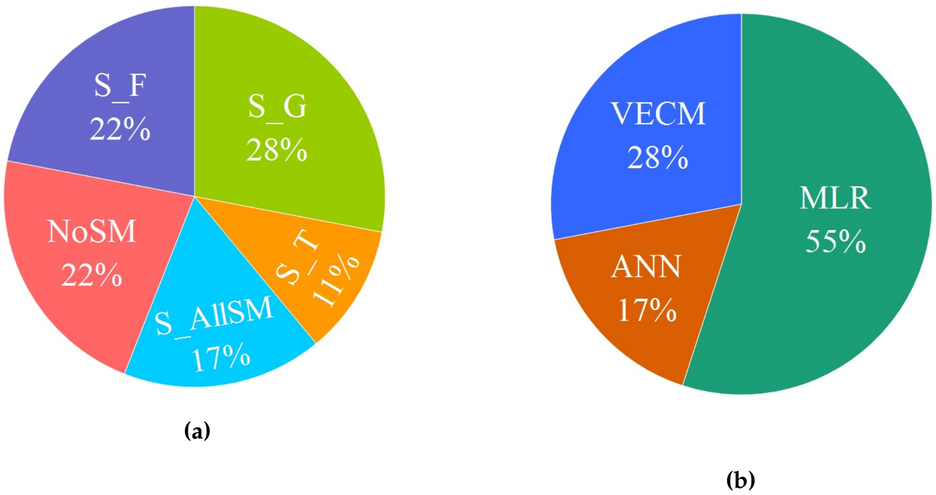
Fig. 6.7.b summarizes the results for MLR and ANN models and different feature groups across all brands based on MAPE. We can see that for the majority of brands Twitter, Facebook, Google and their combination allow to improve the predictions of sales. We can also see that Google are the most helpful features, followed by Facebook and then Twitter (the size of the bars for different feature groups).

Table 6.9: RelMAE results of predicting sales volume using MLR and ANN on the feature set 2. The first line for each brand presents RelMAE results for different features groups, the second line presents relative percentage improvement over the benchmark. Light and dark colors represent improvements by more than 1% and 3%, respectively.

	Multiple Regression					Artificial Neural Networks					Best Model
	S_T	S_F	S_G	S_All	JustSM	S_T	S_F	S_G	S_All	JustSM	
Brand 1	0.98	0.92	0.95	0.98	1.70	1.08	1.00	1.01	1.05	2.54	MLR_F
Brand 10	1.04	1.00	0.94	0.99	1.60	1.29	1.30	1.22	1.49	2.62	MLR_G
Brand 12	0.98	1.00	1.00	1.00	1.22	0.99	1.05	1.03	1.54	8.61	MLR_T
Brand 14	1.10	1.07	0.98	1.14	1.24	1.05	0.95	0.94	0.92	3.94	ANN_All
Brand 15	1.05	1.01	1.02	1.01	1.95	1.13	1.59	1.11	1.66	3.62	MLR_S
Brand 17	1.11	1.10	0.98	1.15	1.09	2.04	2.06	0.98	1.79	8.14	MLR/ANN_G
Brand 18	1.05	1.03	1.01	1.10	1.53	1.28	1.25	1.05	1.15	3.50	MLR_S
Brand 2	0.98	0.97	0.99	1.00	1.50	1.10	1.13	1.27	1.46	3.91	MLR_F
Brand 20	1.08	1.11	1.00	1.07	1.12	0.97	1.62	0.98	1.10	2.73	ANN_T
Brand 21	0.95	0.99	1.01	1.02	1.73	1.12	1.12	1.64	1.14	2.19	MLR_T
Brand 22	1.02	0.94	0.87	0.99	0.92	1.04	1.07	0.99	1.02	1.37	MLR_G
Brand 24	1.01	0.97	0.99	1.01	1.12	0.93	0.89	1.20	0.90	6.85	ANN_F
Brand 3	1.10	1.58	1.05	6.18	1.16	1.10	1.52	1.03	1.68	2.18	MLR_S
Brand 4	1.01	1.03	0.98	1.03	1.24	0.95	1.00	0.97	0.93	6.87	ANN_All
Brand 6	0.92	0.94	0.92	0.92	1.46	1.08	1.08	1.03	1.13	2.42	MLR_T/G/All
Brand 7	1.05	1.04	0.94	0.98	1.64	1.12	1.33	1.07	1.12	3.52	MLR_G
Brand 8	0.99	0.98	0.99	1.00	1.25	1.10	1.42	1.34	1.09	1.58	MLR_F
Brand 9	0.98	0.99	0.89	0.92	2.58	2.13	1.31	2.44	1.85	3.21	MLR_G

Table 6.10: Best model for each brand and a best feature set based on MAPE.

	Benchmark MAPE	Best Model MAPE	MAPE reduction	Relative MAPE reduction	DM test	Best Model Name	Best Feature Set
Brand 1	16.9	15.1	1.8%	10.9%	*	MLR	Facebook
Brand 10	18.2	16.7	1.5%	8.2%	*	MLR	Google
Brand 12	51.0	44.9	6.1%	11.9%	*	VECM	Facebook
Brand 14	12.6	11.2	1.4%	11.1%	*	MLR	Google
Brand 15	21.2	21.2	-	-	-	MLR	No SM
Brand 17	22.7	20.6	2.1%	9.4%	*	VECM	All SM
Brand 18	24.7	22.5	2.2%	9.0%	*	VECM	All SM
Brand 2	15.8	14.9	0.9%	5.5%		MLR	Twitter
Brand 20	23.4	22.9	0.5 %	2.2%	·	VECM	Google
Brand 21	136.1	100.9	35.2%	25.9%	***	ANN	No SM
Brand 22	23.8	21.9	1.9%	8.2%	·	MLR	Facebook
Brand 24	25.2	24.2	1.0%	3.9%	·	MLR	Facebook
Brand 3	18.5	18.5	-	-	-	MLR	No SM
Brand 4	14.0	12.8	1.2%	8.3%		MLR	Google
Brand 6	43.9	36.1	7.7%	17.6%	**	VECM	All SM
Brand 7	7.5	6.7	0.8%	10.7%	*	ANN	No SM
Brand 8	474.0	284.3	189.7%	40.0%	***	ANN	Google
Brand 9	14.2	12.9	1.4%	9.6%	*	MLR	Twitter
***p <0.001,	**p <0.01,	*p <0.05,	and ·p <0.1				



(a)

(b)

Figure 6.10: a) Feature sets of best models; b) Best models by type.

Best Performing Model and Feature Group In this section I choose the best performing model for each brand in terms of MAPE. The results are presented in Table 6.10. From the table we can see that the benchmark outperformed models with social media features only for two out of eighteen brands, brands 15 and 3. The relative MAPE reduction for different brands ranges between 2% and 40%.

To understand how significant is the difference between the benchmark forecasts and ”enhanced” models forecasts I performed a Diebold and Mariano (DM) test. The test statistics, with the stars indicating the level of significance is reported in column DM statistics. The results are based on ”greater” principle, meaning that the forecast of the proposed model should be better than the benchmark. According to the results, 11 out of 18 selected models are significantly different from the benchmark at least at 5% level in out-of-sample forecasting.

Fig. 6.10 summarises the outcomes of best selected models. In 6.10.a we see the distribution of the best feature groups: 28% of best performing models were based on using Google features; Facebook features came second comprising 22% of all models; Twitter was successful in 11% of the cases. It is interesting to notice that when Facebook, Twitter or Google improve sales forecasts individually it is not always the case that the joined effect of those features reduces the overall MAPE (this only happens in 17% of the models).

From 6.10.b we see that the MLR model performed best in 55% of the cases.

The VECM model showed improvement over the MLR for 28% of brands, and ANN was selected as the best model only in 17% of cases.

I also performed a multi-hypothesis Friedman test with a post-hoc Holm correction (Derrac et al., 2011, Demšar, 2006) in order to determine the best model. The results of comparing MAPEs of MLR vs. VECM on Feature Set 1 are presented in Table 6.11. The second column shows the average rank for each model, showing that the VECM with Google features was the best performing model followed by the MLR model with Google features. The performance of the winning model was significantly better than the benchmark (MLR_S) as confirmed by the p-value and Holm correction. The results in Table 6.11 also demonstrate that MLR and VECM models with Facebook features also outperformed the benchmark model, while the model enhanced with Twitter produced the worst results.

Model	Average Rank	Z-score	p-value	Holm Correction
MLR_T	6.4	3.2	<**0.01**	0.006
MLR_All	5.8	2.6	<**0.01**	0.007
VECM_All	5.5	2.267	**0.023**	0.008
VECM_T	5.4	2.2	**0.028**	0.01
MLR_S (benchmark)	5.3	2.067	**0.0396**	0.013
MLR_F	5.1	1.933	0.053	0.0167
VECM_F	4.3	1.067	0.286	0.025
MLR_G	4.1	0.867	0.386	0.05
VECM_G	3.2	-	-	-

Table 6.11: Results of comparing MAPEs of MLR vs. VECM on a Feature Set 1 using Multiple hypothesis Friedman test with a post-hoc Holm correction for $\alpha = 0.05$.

The results of comparing MLR vs. ANN on Feature Set 2 are presented in Table 6.12. Similarly to the results based on Feature Set 1, the MLR models with Google features significantly outperformed the benchmark. We can also see that MLR models enhanced with Facebook and Twitter features were ranked higher in terms of their MAPE performance than the benchmark model trained only on sales data. It is interesting to see that MLR models had a higher rank than ANN models.

6.3.7 Forecasts of Sales Direction

Since prediction of the actual values of future sales is an extremely complicated task I also predict the direction of future sales. The results of accuracy of prediction for Logistic Regression (LR) and ANN are presented in Table 6.13.

Higher values indicate higher accuracies. Values that are improved by more than 1% or more than 3% are highlighted by light and dark colors, respectively.

Model	Average Rank	Z-score	p-value	Holm Correction
ANN_JustSM	11.6	7.488	<**0.01**	0.005
MLR_JustSM	8.8	5.177	<**0.01**	0.005
ANN_All	8.3	4.761	<**0.01**	0.006
ANN_F	8.2	4.669	<**0.01**	0.006
ANN_T	7.3	3.929	<**0.01**	0.007
ANN_G	6.5	3.236	<**0.01**	0.008
ANN_S	5.4	2.334	**0.02**	0.01
MLR_S (benchmark)	5.1	2.080126	**0.038**	0.013
MLR_All	5.1	2.034	**0.042**	0.017
MLR_T	4.7	1.71	0.087	0.025
MLR_F	4.3	1.41	0.159	0.05
MLR_G	2.6	-	-	-

Table 6.12: Results of comparing MAPEs of MLR vs. ANN on a Feature Set 2 using Multiple hypothesis Friedman test with a post-hoc Holm correction for $\alpha = 0.05$.

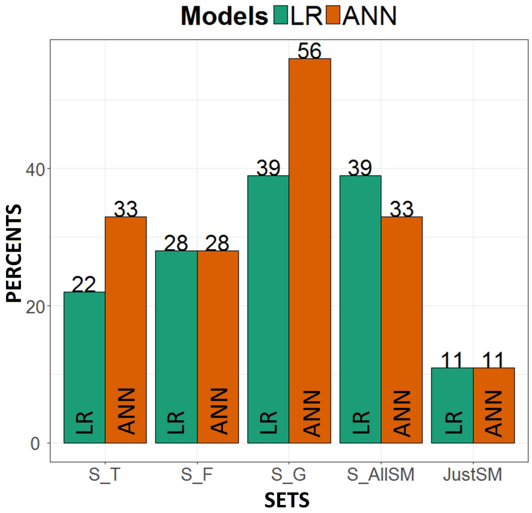


Figure 6.11: Percentage of brands for which performance improved when using social media features compared to the benchmark based on Accuracy metric. Different feature groups are presented along X-axis. Bar colors represent different model types.

For 12 out of 18 brands the accuracy of direction prediction was improved after including social media variables. Fig. 6.11 summarises the results across all brands. We can see that the accuracy was improved more often when Google variables were used, in 39% of brands for LR and in 56% of brands for ANN. For S_T and S_G feature groups ANN allowed to improve predictions for more brands than LR.

The column JustSM in Table 6.13 presents the results of direction prediction for ML and LR using only social media variables. Notably, for 2 out of 18 brands (brands 10 and 9) the models trained only on social media data produced better forecasts than the model trained on historical sales. Along with prediction accuracy

Table 6.13: Results of prediction sales direction using LR and ANN. The first line for each brand presents prediction accuracies for different features groups (%), the second line presents relative improvement over the benchmark. Light and dark colors represent improvements by more than 1% and 3%, respectively.

	Logistic Regression						Artificial Neural Networks						Best
	S	S_T	S_F	S_G	S_All	JustSM	S	S_T	S_F	S_G	S_All	JustSM	Model
Brand 1	68	64 (-5.2)	63 (-6.3)	65 (-3.1)	69 (2.1)	63 (-7.3)	61	64 (4.6)	58 (-5.8)	65 (6.9)	69 (12.6)	61 (-1.2)	LR_All
Brand 10	60	64 (7.1)	68 (12.9)	68 (14.1)	62 (3.5)	63 (5.9)	55	60 (9.0)	62 (12.8)	55 (0.0)	56 (1.3)	56 (1.3)	LR_G
Brand 12	70	63 (-10.1)	61 (-13.1)	66 (-5.1)	51 (-27.3)	60 (-14.1)	67	56 (-16.8)	53 (-21.1)	67 (0.0)	49 (-26.3)	58 (-12.6)	LR_S
Brand 14	77	80 (3.7)	78 (1.8)	76 (-0.9)	77 (0.0)	76 (-0.9)	73	74 (1.9)	72 (-1.0)	74 (1.9)	73 (0.0)	71 (-1.9)	LR_T
Brand 15	73	65 (-11.5)	66 (-9.6)	67 (-8.7)	68 (-6.7)	64 (-12.5)	64	62 (-3.3)	65 (1.1)	67 (4.4)	59 (-7.7)	61 (-5.5)	LR_S
Brand 17	69	69 (0.0)	67 (-3.1)	65 (-6.1)	70 (1.0)	54 (-22.5)	61	65 (7.0)	63 (3.5)	63 (3.5)	62 (2.3)	51 (-15.1)	LR_All
Brand 18	69	68 (-2.0)	65 (-5.1)	67 (-3.1)	65 (-5.1)	68 (-2.0)	68	64 (-6.2)	66 (-3.1)	61 (-10.3)	63 (-7.2)	59 (-13.4)	LR_S
Brand 2	65	67 (3.3)	68 (4.4)	63 (-2.2)	64 (-1.1)	62 (-4.4)	61	60 (-2.3)	61 (-1.2)	63 (2.3)	65 (5.8)	59 (-3.5)	LR_F
Brand 20	63	66 (4.4)	70 (11.1)	69 (8.9)	70 (11.1)	58 (-8.9)	67	68 (2.1)	61 (-8.4)	70 (5.3)	63 (-6.3)	57 (-14.7)	LR_F
Brand 21	75	70 (-6.6)	72 (-3.8)	73 (-1.9)	72 (-3.8)	62 (-17.0)	65	65 (0.0)	64 (-2.2)	68 (4.3)	68 (3.2)	58 (-10.8)	LR_S
Brand 22	88	85 (-4.0)	87 (-0.8)	84 (-4.8)	88 (0.0)	77 (-12.0)	78	78 (0.0)	76 (-2.7)	75 (-4.5)	74 (-5.4)	70 (-10.8)	LR_S
Brand 24	56	56 (0.0)	67 (18.8)	61 (8.8)	60 (6.3)	55 (-2.5)	58	55 (-4.9)	59 (2.4)	53 (-8.5)	56 (-2.4)	50 (-13.4)	LR_F
Brand 3	69	68 (-1.0)	58 (-16.3)	70 (2.0)	62 (-10.2)	54 (-21.4)	63	59 (-5.6)	61 (-2.3)	63 (0.0)	58 (-7.9)	51 (-19.1)	LR_G
Brand 4	76	73 (-3.7)	74 (-2.8)	75 (-1.9)	73 (-4.6)	70 (-8.3)	70	67 (-5.0)	69 (-2.0)	73 (4.0)	70 (0.0)	61 (-14.0)	LR_S
Brand 6	63	61 (-3.4)	60 (-4.5)	65 (3.4)	58 (-6.7)	54 (-13.5)	63	51 (-18.9)	56 (-12.2)	58 (-7.8)	49 (-23.3)	54 (-14.4)	LR_G
Brand 7	80	81 (0.9)	80 (0.0)	82 (2.6)	80 (-0.9)	80 (-0.9)	73	69 (-4.9)	73 (1.0)	77 (6.8)	69 (-4.9)	68 (-6.8)	LR_G
Brand 8	65	65 (0.0)	58 (-10.8)	65 (0.0)	66 (1.1)	59 (-9.7)	69	59 (-14.3)	59 (-14.3)	62 (-10.2)	59 (-14.3)	59 (-14.3)	LR_All
Brand 9	62	61 (-1.1)	61 (-1.1)	64 (3.4)	63 (2.3)	63 (1.1)	54	61 (11.7)	61 (11.7)	63 (16.7)	56 (3.7)	65 (20.1)	ANN_SM
Mean	69	68	68	69	68	63	65	63	63	66	62	59	

I calculated a 95% binomial confidence interval, which ranges between 5% and 7%. It is important to notice, that the 95% confidence interval for JustSM accuracy for almost all brands (apart from brands 17, 24, 3 and 6) is above 50%, to be more precise, on average 63.4% ±6.7 for LR and 59.4% ±6.8 for ANN. Since the whole interval is higher than 50% for the majority of brands I conclude that Twitter, Facebook and Google Trends contain information about future sales, moreover, social media data without sales information can be used to predict the direction of next day sales of a brand.

6.4 Chapter Conclusions

In this chapter I investigated whether information from social networks, such as Facebook and Twitter, and search data from Google Trends have the ability to improve daily sales forecasts. The forecasting performance of linear and non-linear forecasting models was compared through a case study of 18 world-known brands. As a general result across brands, I found that incorporation of Twitter, Facebook and Google Trends data into models leads to improvements of forecasts compared to the autoregressive model of past sales only, thus, highlighting the ability of consumer sentiment and brand perception to effect future sales. Moreover, I identified that social media data alone, without sales information, can be used to predict sales direction with 63% accuracy. I also found that imposing cointegrating restrictions between sales and social media variables and using Vector Error Correction Model yields a relative forecasting improvement of MAPE ranging from 8% to 17% versus a simple autoregressive model.

The results of the findings are listed as answers to the research questions stated at the beginning of the chapter, and can be summarised as follows:

1. Models that had at least one of the social media factors included, outperformed models trained only on past sales data in 78% of the cases for predicting sales volume and in 67% of brands for predicting next day sales direction. This result indicates that social media contains information about future sales and management of companies should consider collecting Twitter, Facebook and Google data and including it into their forecasting models. The results also showed that it is possible to predict next day sales

direction with the accuracy of 63% using only social media features. This means that the change in volume and sentiment of online chatter during the past few days is a direct indicator of future sales direction.

2. MAPE results across brands reveal that approximately 28% of the best models contained Google features; 22% contained Facebook features; 11% contained Twitter features and 17% included the combination of all three types of social media variables. Google Trends variables consistently appeared to be indicators of future sales allowing to improve forecasts for most of the brands across different feature sets and different models, while Twitter features were successful in reducing MAPE just for a few brands. It is interesting to observe that when at least one social media indicator helps to improve forecasts for a particular brand, other social indicators also have an impact. For example, for brands 1, 2, 6, 7, 8, 10, 21, 22 and 24 at least two social media variables helped to improve forecasts (as judged from Table 6.9, MLR part). This might be related to the fact that if a brand is engaged in social media marketing and building the relationship with its fans online, it is likely to do so across all types of media.

3. Johansen test revealed that there is a cointegrating relationship between sales, Twitter and Facebook for all analysed brands, and between sales and Google Trends for some brands. Incorporation of the long-run restriction through an error correction model yielded superior forecasts in comparison with first difference model in 40% of cases when Twitter data was considered; in 53% when Facebook variable were included; in 67% when Google features were added and in 40% of cases when a combination of social media variables was included in the model. Among the best models across all brands, the VECM performed better than other models in 28% of cases. The cointegrating relationship between sales and consumers' response (expressed through Twitter sentiment, Facebook likes, volume of Google searches) means that as people feel more positive about some products, they are more likely to buy those products. In other words, there is a long-run equilibrium relationship between sales volume and consumers' social response. If negative opinion about a company or product spreads through the web, the sentiment of social

media will decrease. The effect of how quickly or slowly sales would drop in response to negative news can be estimated using VECM. In this experiment we observe that the estimated lag between changes in consumers' social response and reaction in sales is, on average, 7 days for Twitter, 9 days for Facebook and 14 days for Google.

4. ANNs outperformed MLR and VECM only in 17% of cases across all brands, however, the performance of ANN should be investigated further on a larger dataset, since current dataset lacked a sufficiently large number of observations to avoid over-fitting.

Chapter 7

Conclusions and Future Work

This chapter presents the main findings of the thesis. The limitations of the research are also discussed followed by suggestions of possible improvements and suggestions for further development of this research topic.

7.1 Conclusions

Sales forecasting is an important topic in many business and economic settings. Previous research have shown that by understanding consumer demand and by producing better sales forecasts for products, a company can improve the efficiency of its supply chain, and, thus, gain competitive advantage over its rivals. In the context of growing impact of online social networks, a question arises, whether information from online user-generated data be used by manufacturers to understand and forecast product demands better?

In this thesis I aimed to answer this question by conducting a large scale study of 75 brands from the retail sector, including apparel, footwear, accessories, body care, furnishings, equipment and games. Differently from the previous literature, I measured the impact of social media on sales utilising two different approaches: 1) by analysing the power of Twitter events to predict spikes in sales; 2) by comparing sales forecasts from models trained on historical sales data only and models trained on sales data combined with Twitter, Facebook and Google Trends information. The findings and conclusions obtained by two approaches are summarised below:

- **Events Analysis.** The majority of existing works, that studied the effects of social media on sales, used regression models to perform explanatory

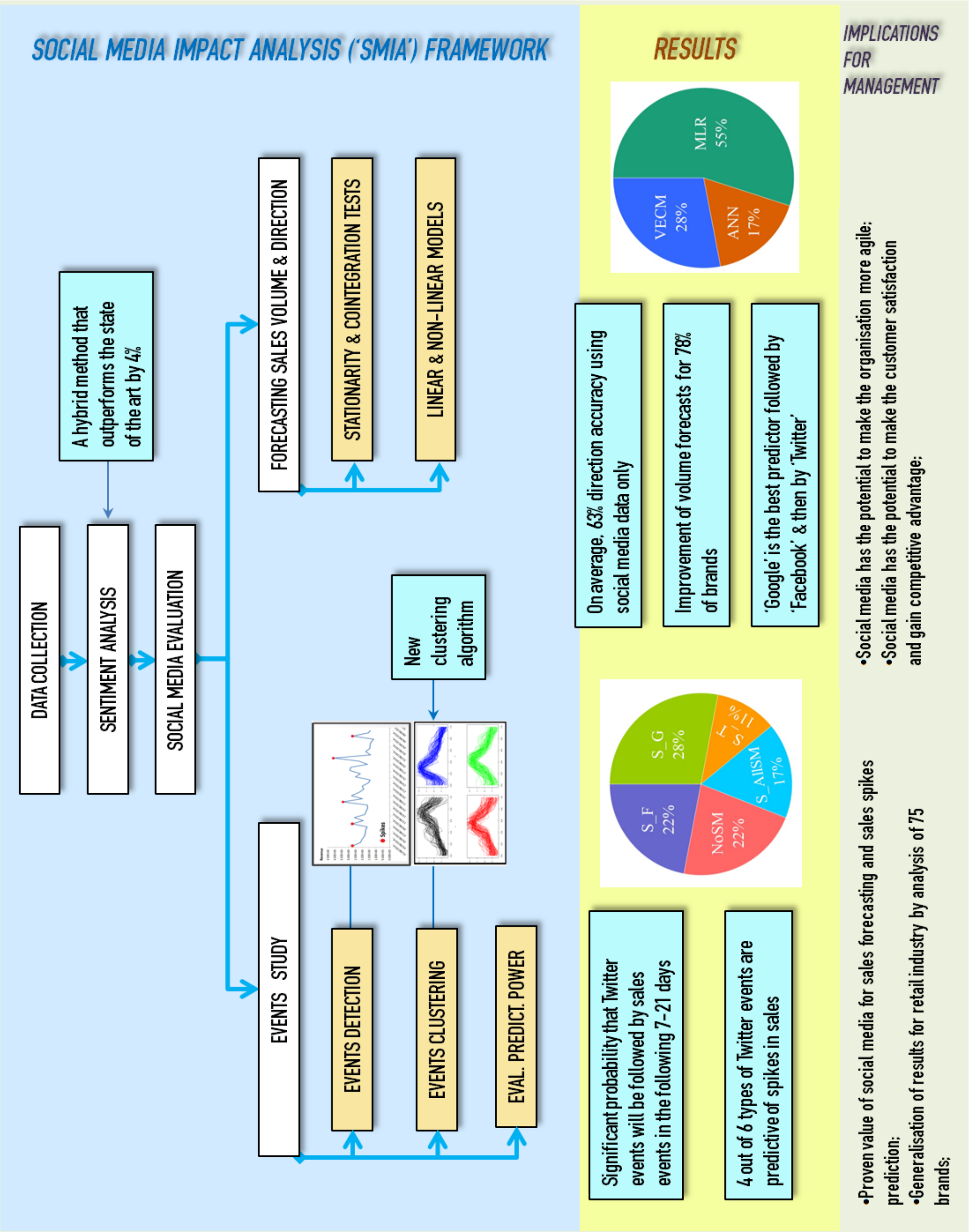

SOCIAL MEDIA IMPACT ANALYSIS ('SMIA') FRAMEWORK
DATA COLLECTION
SENTIMENT ANALYSIS
SOCIAL MEDIA EVALUATION
A hybrid method that outperforms the state of the art by 4%
FORECASTING SALES VOLUME & DIRECTION
STATIONARITY & COINTEGRATION TESTS
LINEAR & NON-LINEAR MODELS
EVENTS STUDY
EVENTS DETECTION
EVENTS CLUSTERING
EVAL. PREDICT. POWER
New clustering algorithm
Spikes
RESULTS
MLR 55%
VECM 28%
ANN 17%
On average, 63% direction accuracy using social media data only
Improvement of volume forecasts for 78% of brands
'Google' is the best predictor followed by 'Facebook' & then by 'Twitter'
S_G 28%
S_F 22%
NoSM 22%
S_AllSM 17%
S_T 11%
Significant probability that Twitter events will be followed by sales events in the following 7-21 days
4 out of 6 types of Twitter events are predictive of spikes in sales
IMPLICATIONS FOR MANAGEMENT
•Social media has the potential to make the organisation more agile;
•Social media has the potential to make the customer satisfaction and gain competitive advantage;
•Proven value of social media for sales forecasting and sales spikes prediction;
•Generalisation of results for retail industry by analysis of 75 brands;

Figure 7.1: Schematic representation of the main thesis conclusions.

or forecasting analysis of social media variables. This thesis presents a contribution to the field by proposing to study events in Twitter and measure their sales predictive power (Chapter 5).

As part of this research, a framework for data collection, sentiment analysis, events detection and events clustering was developed (Figs. 3.1, 4.1, 5.2).

During the first step, I performed the analysis of all Twitter events without clustering them into different groups. The results of analysis indicated that when a spike in Twitter sentiment is observed for a particular brand, there is a significant probability to observe a spike in sales for that brand in the following 3 weeks (see Fig. 5.15). I investigated the reasons behind some of the spikes in Twitter using text analytics (Appendix A, Tables 8.1 - 8.5) and discovered that spikes can occur due to: a) external events with strong brand association (exogenous); b) internal brand events (endogenous); c) social chatter driven by the brand (exogenous); d) social chatter driven by people (endogenous). To study the predictive power of different events' dynamics, as the second step, I performed clustering of events based on the slopes of their growth and relaxation signatures. Through the clustering process distinctive shapes of Twitter events were detected (see Figs. 5.10, 5.13): a) events with symmetric growth and relaxation signatures; b) events with a long growth signature and a short relaxation signature; c) events with a short growth signature and a long relaxation signature. Within each group a variation of events' width was observed resulting in six clusters of events. By analysing the predictive power of each Twitter event cluster I identified that different types of Twitter events appear before sales events in a non-random manner (Tables 5.2 and 5.3). Moreover, four clusters of events (types 2, 4, 5 and 6, Fig. 5.17) have a significant probability to be followed by spikes in sales within the following 2 - 3 weeks, and that probability is higher than the one calculated for non-clustered Twitter signal. This result demonstrates that events clustering can be used as part of signal filtering, allowing to distinguish types of social media events that have significant predictive power.

- **Forecasting Sales Volume and Direction.** Following the traditional approach, I also performed forecasting of sales volume and direction. However, differently from previous works, I compared the impact of three types of online data, Twitter, Facebook, Google Trends and performed short-term, rather than long-term forecasting.

  According to the results of feature selection process, multiple social media variables were selected among the 30 most important features for all brands (see Fig. 6.4.a), indicating that Twitter, Facebook and Google Trends variables have the ability to explain future sales. When comparing the performance of models with and without social media indicators based on MAPE, the models that had social media information included outperformed the benchmark model in 78% of cases (see Table 6.10 and Fig. 6.10). The relative improvement of MAPE ranged between 2% and 40% and was confirmed to be significant based on the Diebold and Mariano test.

In terms of the best performing indicators, Google Trends showed to be the most consistent predictor of sales volume: it allowed to improve forecasts for 67% of brands (Fig. 6.7.a) and the improvement was the largest compared to other indicators in 28% of cases (Fig. 6.10.a). A better performance of Google Trends compared to Facebook and Twitter might be related to the fact that people do Google search when they have a strong intention to buy something, while the discussion on social networks might happen without a purpose to purchase, as a response to other people's comments or as a reaction to messages posted by a brand itself. The effect of Twitter and Facebook on sales performance showed to be rather complex and differed from brand to brand. Models that included Facebook features improved sales forecasts for 53% of brands and outperformed other models in 22% of cases; for Twitter these numbers were 27% and 11%, respectively (see Figs. 6.7.a and 6.10.a). To understand the key factors which lead to Facebook and Twitter to be predictive of sales one should dive deeper into brand's social media strategies. Some companies might be very active in promoting new products and responding to their consumers' sentiment, while other brands might still have a very low level of engagement with its followers. Overall, this means that simply adding Twitter or Facebook information to a model will not necessarily improve its accuracy. The relevance of including this information should be considered on a case by case basis.

Very positive results were achieved for all social media data sources when predicting the direction of sales. The accuracy of direction prediction was improved for 67% of brands when at least one type of social media variables was included. More importantly, for 14 out of 18 brands the direction of sales was predicted with the average accuracy of 63% even when historical sales data was not part of the model. These findings demonstrate that by observing the change in opinion through online user-generated content companies can predict the change in sales direction within just a few days, and take the necessary actions to adjust the supply.

In terms of models performances, the VECM performed better than other models in 28% of cases, reducing the relative MAPE by 8% to 17% and demonstrating the importance of considering the cointegrating relationships between variables. ANNs outperformed linear models only for 17% of brands, however, the performance of ANNs should be investigated further with bigger datasets.

To summarise, as a general result across brands, the findings of these thesis demonstrated that Twitter, Facebook and Google Trends data allows to achieve improvements of sales volume forecasts. Moreover, the direction of sales can be predicted based on social media data only, and events in social media can be used to predict spikes in sales. These results have serious implications for both, academia and practitioners:

- **Implications for Academia.** This thesis contributes to the field of social media analysis by providing empirical evidence that consumer sentiment expressed online effects future sales. For the first time a large scale analysis of 75 brands was conducted for ”non-trending” products, generalising the findings for retail industry. The outcomes of this research also allow to generalise the conclusions across multiple online data sources, Twitter, Facebook and Google Trends.

  Through research process a few theoretical contributions were made: 1) a hybrid method for sentiment analysis was developed, combining the benefits of the lexicon-based approach and the machine learning approach; 2) a new events clustering method and an approach for automatic detection of the optimum event window were proposed, extending the event study field.

  Moreover, I suggested a ’SMIA’ framework that contains all the steps necessary for evaluation of social media impact: data collection, sentiment analysis, feature engineering, feature selection, non-stationarity testing, cointegration testing, and possibility for training different types of models: multiple linear regression, artificial neural networks and vector error correction model. This framework can be directly adopted by researchers who seek to measure the predictive power of social media in different research settings.

- **Implications for Industry.**

  This study also provides practical implications for business managers. Previous research has illustrated that investment by companies in social media provides a strong return on investment in long term (Hudson and Thal, 2013). However, the evidence of the impact of social media in short term, such as relationship with next-day sales, is very scarce. The research at hand fills this gap and provides empirical confirmation that social media information can be used to predict next-day sales and, in many cases, predict sales spikes. Based on these findings, companies should reconsider their social media strategies in order to have the capacity to develop day-to-day communication with their online communities, monitor the level of customer satisfaction daily and respond to consumer requests appropriately and timely. The managers could increase the customer satisfaction by delivering positive experiences on a regular basis, for example, by providing discounts, organising events, offering gifts to the most loyal customers. According to Brandwatch[1] 96% of the people that discuss brands on social media do not follow those brands' official profiles. This means that in order for companies to "stay on the pulse" of what is happening online and truly understand their customers' needs, companies should go beyond their own channels and monitor all conversations online related to their brand. Such approach requires a strategy and development of a framework for integrating the social media at every step of the process. The framework for social media analysis proposed in this thesis is an attempt to illustrate the practical value of

[1] Brandwatch provides insights to brands based on social media data: https://www.brandwatch.com/

bringing together behavioural theories, models, and latest technologies. This framework creates a basis for the design and development of social media applications, and can help companies and system developers to formulate and implement their own social media application strategies.

To conclude, it is important to mention that before a firm gets into the social media arena, it should define the appropriate objectives it wants to achieve through social media engagement. These objectives should consider not just the firm, but also benefit the customer, since social media is predominantly about engagement and collaboration. If planned well, social media has the potential to transform the organisation by making it more agile and able to respond quickly to changing consumer demand. Agility in the market place normally equals sales, higher revenue and competitive advantage.

7.2 Limitations and Future Work

This study has several limitations that need to be addressed in future research. The first limitation is related to data quality. While social media data presents a great source of information for academics and practitioners, it is also important to acknowledge that a lot of this data might be erroneous or irrelevant. The popularity of social networks made them particularly attractive for spammers, who spread advertisements and viruses (Zheng et al., 2015, Zhu et al., 2012), phishers, who aim to attend sensitive users' information, or individuals, who pollute the content by spreading rumours or trying to compromise brands' reputations (Benevenuto et al., 2010). Another issue that effects data quality is the keyword ambiguity. The retrieval of social media data is usually performed by specifying search keywords. If a keyword, for example, a brand name, is also a commonly used word, such as "Apple", the returned data will contain a lot of irrelevant information. Therefore, a major challenge when analysing social media data is to separate meaningless messages from the relevant content that could contain valuable information. Data filtering performed in this thesis is limited to minimizing the ambiguity of search keywords by selecting only the brands that have distinct names, for example, Adidas. Procedures, such as spam detection or rumours detection were not performed at this stage of research, because traditional text analysis techniques are not suitable for short social media messages that contain a lot of grammatical and spelling errors, mixed languages, slang and abbreviated words. Addressing spam filtering problem presents a challenging research topic in its own right and was not the objective of this thesis. However, improving data quality is an essential requirement for producing reliable results and should be performed during the future work on this research topic.

Another limitation is related to the aggregation level of data. Sales time series analysed in this study were generated by aggregating the revenues across different products categories on a global scale. While this approach allows to reflect the overall sentiment of online community towards a brand, predicting performance of individual products in different geographies would require to perform product based aggregation and taking into account spatial information. Evaluating the role of social media data for product sales presents an objective for the future research.

The events study performed in this research is limited to identifying different types of social media events based on their shapes. However, understanding the meaning/reasons behind different events could shed even more light on consumers' intentions and different mechanisms of information dissemination. One of the main directions for future research is to interpret what different events' shapes represent in terms of social media dynamics (persistence of news, importance, endogeneity/exogeneity, etc.). To address this problem one could use topic detection techniques in order to understand the content of conversations during the spikes. Incorporation of information about marketing campaigns would also be extremely useful for identifying whether a particular type of spike accrues due to promotions or as a result of natural consumer interest to a product. Additionally, by combining traditional features for events clustering, such as textual, spatial, temporal, and network information with the shape based features proposed in this thesis, one could obtain even more meaningful clusters. As the next stage of future work it would be beneficial to incorporate the extracted knowledge about different event types into forecasting models for consumer sales. In terms of models, methods used in this thesis were limited to Multiple Linear Regression, Vector Error Correction Model and Artificial Neural Network. As it has been mentioned earlier, the performance of ANN was not conclusive due to the limited number of data points available for each brand. To further investigate the non-linear relationships between sales and social media indicators more data should be obtained. As part of future research, it is promising to also evaluate the performance of Extreme Learning Machine, a neural network that learns much faster than traditional feedforward NN, and a Long Sort-Term Memory algorithm, a recurrent neural network specifically designed for time series prediction.

One more constraint that I would like to acknowledge, is that data collected and analysed in this thesis is limited to textual information (Twitter messages, Facebook posts and comments). During the last three years, visual information, such as images and videos, became an essential part of social media content. Some social media experts would say that we shifted from social media to visual social media (Moritz, 2016). People use images and video to express their preferences, and companies use visual content to communicate more easily with their followers. According to Cisco, by 2019 video content will make up 80% of all internet traffic (Guta, 2016). In this context, a combined analysis of textual and visual information becomes a necessity. Future studies in this area should generalize the findings in social media impact analysis by including Youtube, Instagram and Pinterest platforms in the pipelines for data collection.

Finally, as the future development, the 'SMIA' framework proposed in this thesis can be further developed not only for the purpose of sales forecasting, but for many other applications, such as accurate consumer profiling, new products design and price planning.

Chapter 8

Appendices

A. Twitter Events Exploratory Analysis

The first step of time series analysis is displaying the time series and trying to drive any insights from the basic graphs. While performing such analysis, I identified the presence of events (outbursts of activity) in sales and social media time series. In this section, I performed exploration of potential reasons for such spikes using as an example time series of Twitter sentiment data for some of the brands. The analysis of spikes was performed using the following steps:

1. Identifying the time frame when the spike happened.

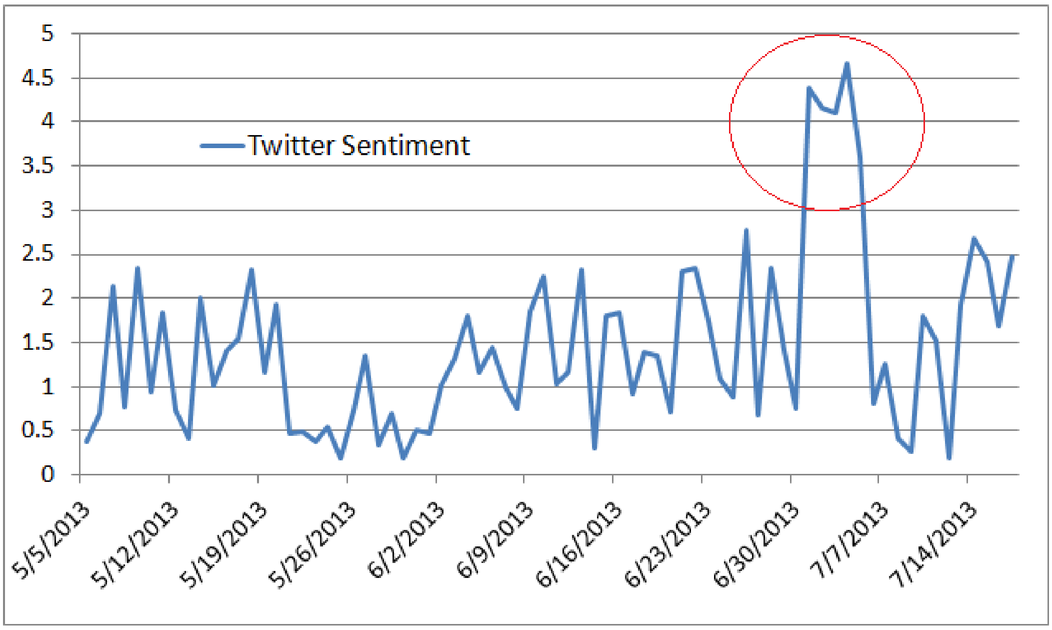
2. Identifying the most popular tweets and re-tweets by performing text analysis.

3. Analysing the content of the most popular re-tweets in order to identify the potential reason for a spike.

4. Using Google search to identify potential events related to the brand during the ”spike period”.

Tables 8.1 - 8.5 provide examples of sentiment spikes for some brands with identified reasons. For the purpose of data protection I do not mention brands' names, instead, I use numerical identification in order to distinguish between different brands.

Brand1



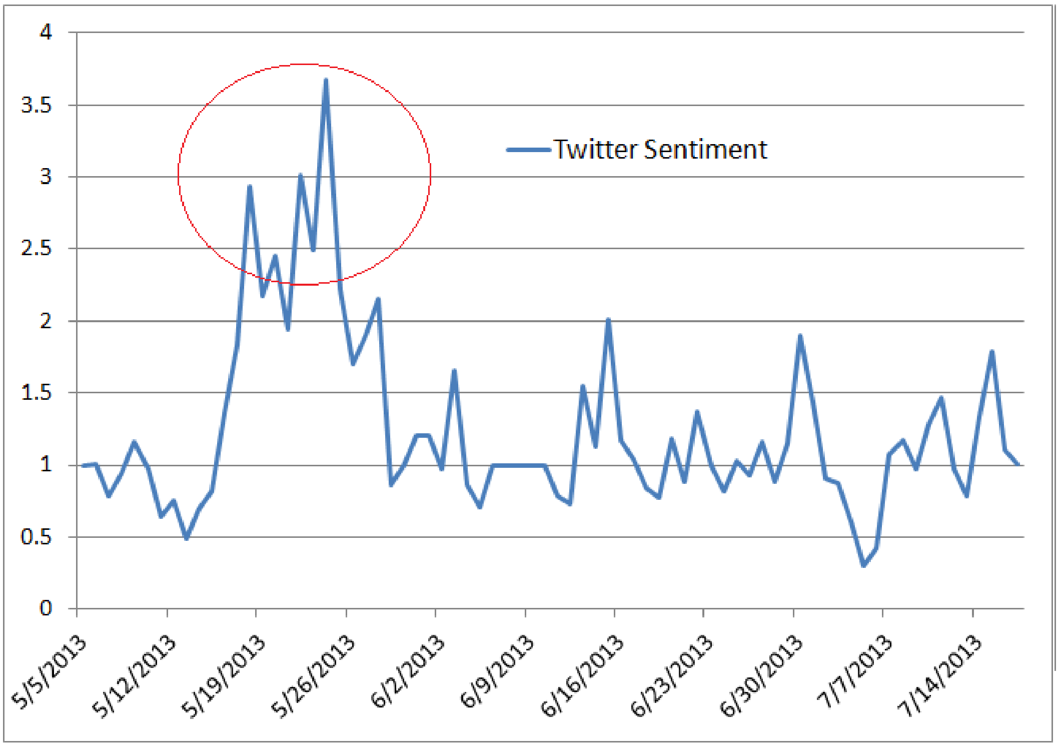
Event dates: 30.06.2013 - 07.07.2013
Description:

- A spike in positive sentiment.
- On average, the number of tweets per day was not exceeding 500 tweets.
- During the event more than 8,000 tweets were posted and the positive sentiment ratio of messages was about twice higher than usual.
- Almost all of the messages were re-tweets of the following tweet: ”RT BringingAdvert: #fashion #new #best #seller Have a look at the latest ”trend” jeans!! `http://t.co/mFLdlxEBSd` #news #CNN #BBC #Brand1”.

The reason:
The spike was related to the release of a new jeans collection. Originally, the message was posted by BringingAdvert, an advertising company. The message caused interest of the wide range of people and got re-tweeted by them. This demonstrates, that the information about new product lines can be spread using Twitter and reach wide audiences. In this sense, Twitter can help brands to increase brand awareness.

Table 8.1: Brand1 event description.

Brand2



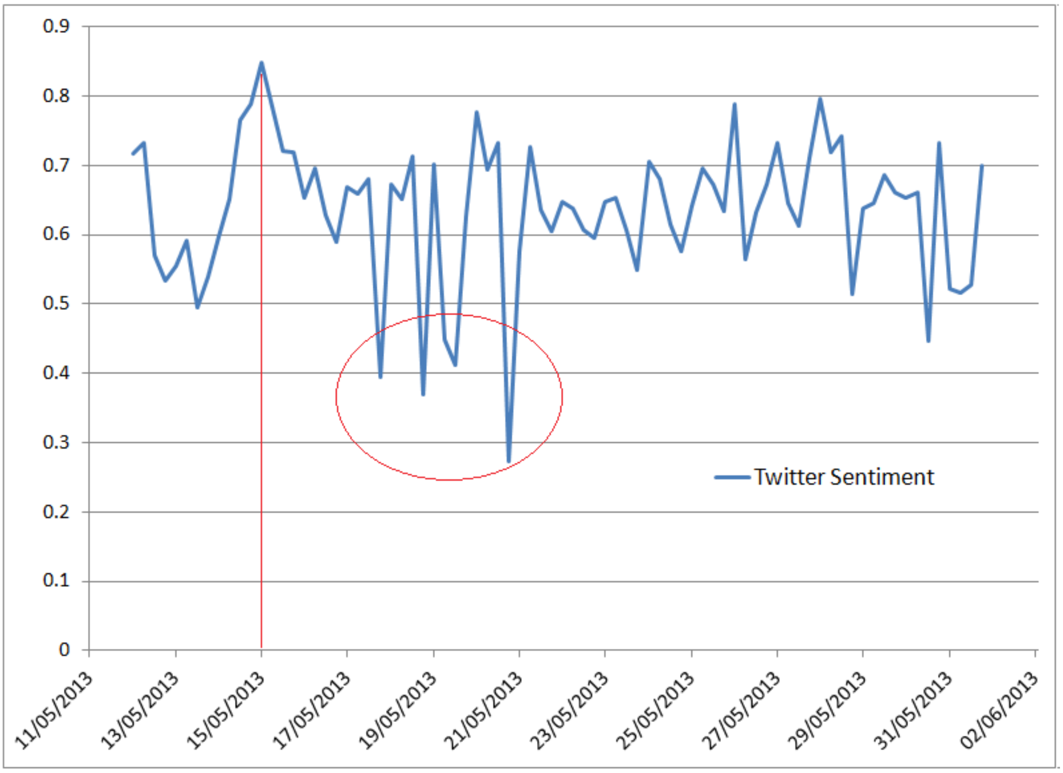
Event dates: 14.05.2013 - 29.05.2013
Description:

- A spike in positive sentiment;
- During the event about 38,000 tweets in English were published.
- 22,514 tweets among 38,000 were re-tweets.
- Almost 15,000 messages contained words ”League” or ”Soccer”.
- Half of the re-tweets contained the following message: ”RT SoccerBible: If Bayern Munich win the Champions League we'll give away a pair of Brand2 Predators! Follow & RT to enter. `http://www.nottinghamforest.co.uk/news/article/forest-team-up-with-Brand2-878245.aspx`”.

The reason:
25th of May 2013 was the day of the UEFA Champions League Final. Brand2 was one of the sponsors of the event. Few days before and after people were discussing the event and were re-tweeting the message regarding the pair of Brand2's shoes that was given away as a present in case Bayern Munich won. This spike demonstrates, that major social or cultural events are reflected on Twitter through the increase in Tweets volume and change of sentiment. It also shows that encouragement of brands to re-tweet their messages can help to increase brand awareness.

Table 8.2: Brand2 event description.

Brand3



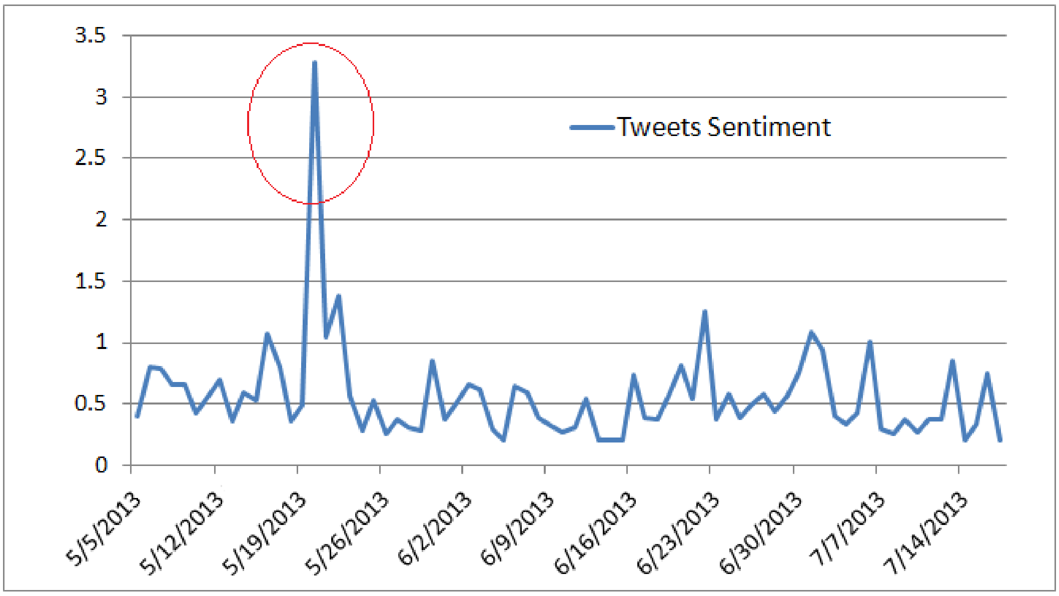
Event dates: 15.05.2013 - 21.05.2013
Description:

- Multiple draw-downs of sentiment ratio;
- Sentiment dropped below 0.5 multiple times during the event, indicating that more negative tweets during each day were posted than there were positive tweets.
- The value of sentiment ratio below 0.5 is unusual since on average the volume of positive tweets is higher than than the volume of negative tweets.

The reason:
Google search for possible reasons of sentiment change revealed that on the 15th of May a video was published online, in which a man was distributing Brand3's clothes to homeless people. The video caused a resonance in public and a few articles were published next day, recalling the speech of Brand3's CEO that he gave months ago and which did not receive social approval. In that speech the CEO highlighted that Brand3 was exclusive for a specific type of people only. During the next days many more articles were published online related to Brand3's CEO speech, causing new waves of negative response on Twitter and other social media. The analysis performed few months after the event revealed that sales for the brand dropped dramatically and the CEO had to step down from his position. After the resignation of the CEO shares of the company went up. This event demonstrates, that Twitter sentiment score is a strong indicator of peoples attitude towards brand's actions and brand's values.

Table 8.3: Brand3 event description.

Brand4



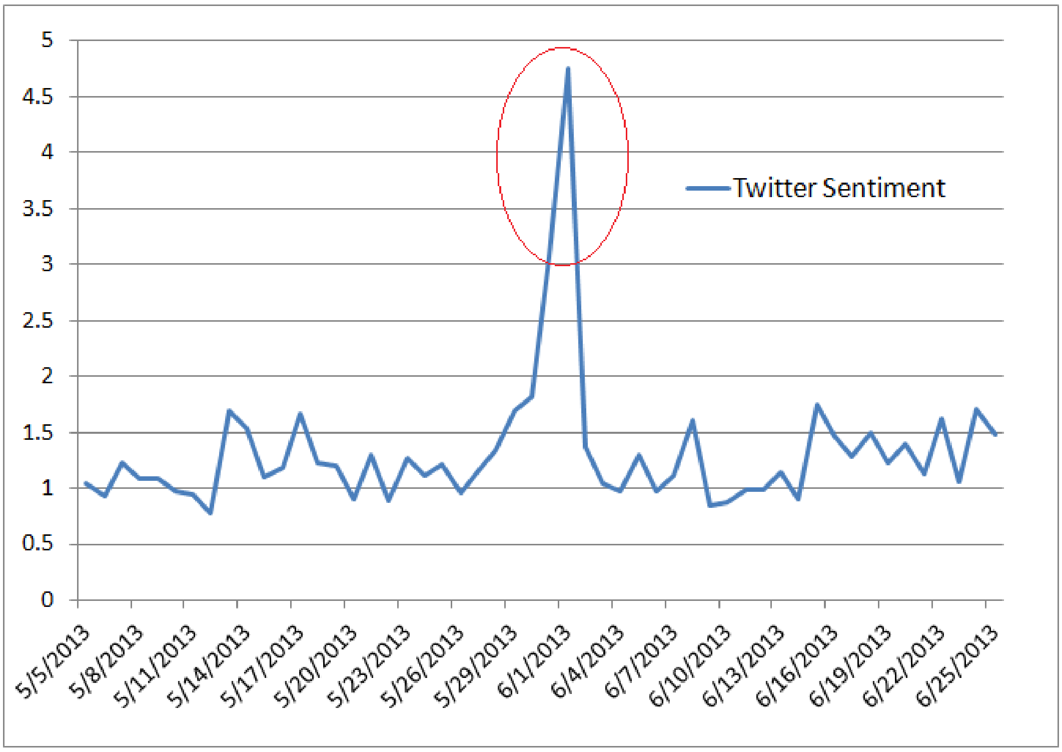
Event dates: 05.07.2013 - 10.07.2013
Description:

- Brand4 is a Japanese fashion retailer.
- A spike in positive sentiment was detected.
- 2259 tweets about Brand4 were posted during the event period.
- 2007 of messages were re-tweets.
- The most popular re-tweets included URL links to a funny T-shirt and an apron (see picture on the right)

The reason:
The spike in sentiment was a result of people actively re-tweeting links to some of the products offered by Brand4. This spike seems to have an endogenous nature and demonstrates that Twitter can serve as a platform for genuine discussions of brand's products by consumers. The content and sentiment of these discussions can be used by brands to make further decisions regarding its product lines.

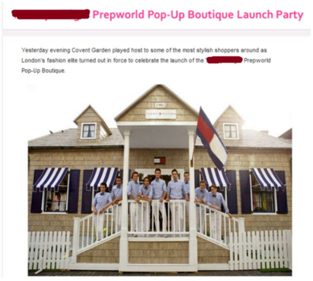
Table 8.4: Brand4 event description.

Brand5



Event dates: 28.05.2013 - 04.06.2013
Description:

- A spike in positive sentiment.
- During the spike in positive sentiment, on average 3000 tweets were posted per day, which is about 70% more tweets than the average amount of tweets for Brand4 per day.
- From those tweets about one third of messages mentioned ”Launching Party”.



[redacted] Prepworld Pop-Up Boutique Launch Party

Yesterday evening Covent Garden played host to some of the most stylish shoppers around as London's fashion elite turned out in force to celebrate the launch of the [redacted] Prepworld Pop-Up Boutique.

The reason:
On the 1st of May 2013 Brand5 opened a new boutique in Soho and hosted a launching party. The spike in Twitter data demonstrated that big events, in which a brand participates, get reflected through the volume and sentiment of Twitter messages.

Table 8.5: Brand5 event description.

Conclusions. The analysis of some events in Twitter signal demonstrated that Twitter sentiment and Twitter volume allow to capture people's perceptions of the brand, analyse reactions to brand's activities and determine an aggregated opinion about new products. Overall, four distinct types of event were discovered (it is import to keep in mind that further investigation might uncover more dynamics):

- External events with strong brand association. For example, brand participating in sponsoring of large sport events (Brand2 example). This event type is described by high Twitter activity that does not necessarily result in sales increase. However, the event contributes to increase in brand awareness.

- Internal brand events. Examples include opening a new store (Brand5) or expressing brand's values and future goals (Brand3). This event type receives strong reflection through Twitter sentiment in conjunction with the tweets volume. It is expected to see an impact on sales as well, but still to be proven with additional sales data.

- Social chatter driven by the brand, such as a new product line release (Brand1), sales, promotions, marketing campaigns. The spike in twitter volume or sentiment is likely to be followed by a spike in sales. The opposite can also be true, when a spike in sales is followed by an increase in Twitter volume and sentiment.

- Social chatter driven by people (Brand4). This event type is described by high volume of Twitter discussions that related to some of brand's products and that were not initiated by the brand. The event in Twitter might be preceded or followed by a spike in sales and provides an opportunity for the brand to evaluate people's opinions about some of the products. This information can be used to adjust value proposition or produce more accurate sales forecasts.

B. Clustering of Benchmark Datasets

Figs. 8.1- 8.3 present the results of clustering for the benchmark datasets: mallet dataset, synthetic dataset and the wheat dataset. Since all time series in datasets are of equal length, I only compared the performance of hierarchical clustering based on Euclidean Distance and based on Euclidean Slopes Distance. The clustering results for mallet and synthetic datasets produced accuracy of 100% for both methods (Fig. 8.1, Fig. 8.2). The clustering results for the wheat dataset using

Euclidean Slopes Distance approach were higher (78%, Fig. 8.3) than the results achieved based on Euclidean Distance (56%, Fig. 8.4).

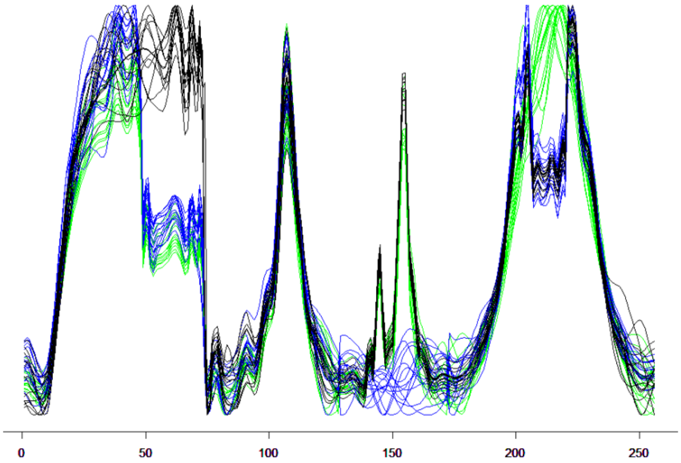


Figure 8.1: Clustering results for the mallet dataset. Both approaches, hierarchical clustering based on Euclidean distance and based on Slopes based Euclidean distance produced accuracy of 100%.

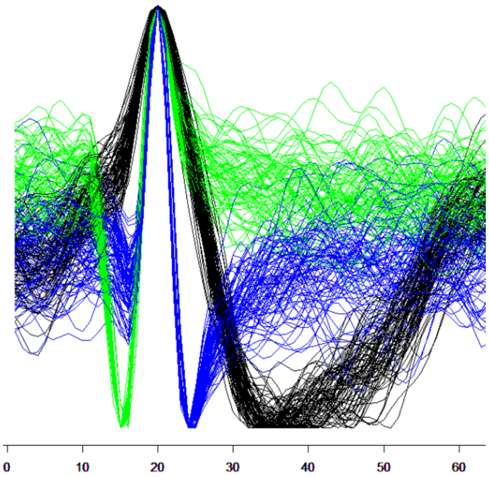


Figure 8.2: Clustering results for the synthetic dataset. Both approaches, hierarchical clustering based on Euclidean Distance and based on Euclidean Slopes Distance produced accuracy of 100%.

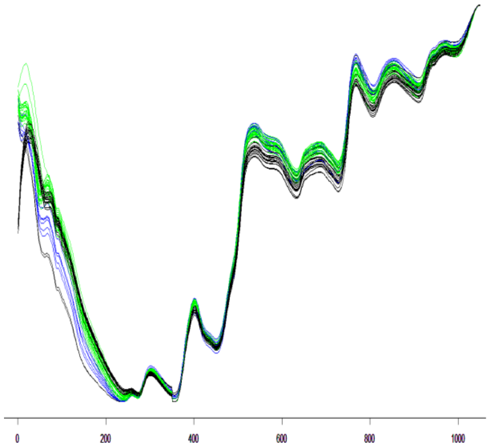


Figure 8.3: Clustering results for the wheat dataset for hierarchical clustering based on Euclidean Slopes Distance produced accuracy of 78%.

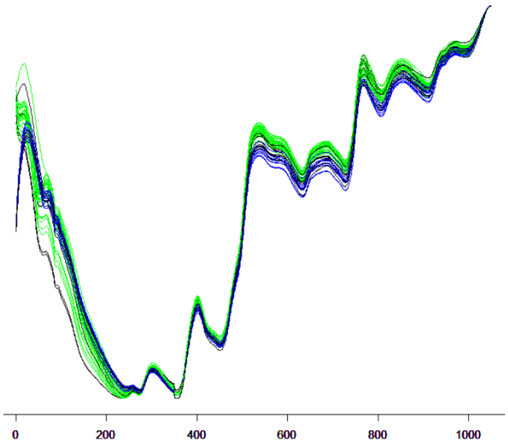


Figure 8.4: Clustering results for the wheat dataset for hierarchical clustering based on Euclidean distance produced accuracy of 56%.

C. Histograms of Proportions of Social Media Features Selected for Different Brands

The proportions of social media features selected for each feature group for each brand are presented in Figs. 8.5-8.8. We see that for all brands social media features were selected among the important. For example, for brand 5, group *S_T* (Fig. 8.5, 40% of features were Twitter related; for brand 11, group *S_AllSM* (Fig. 8.8, 50% of features were a mix of different social media features.

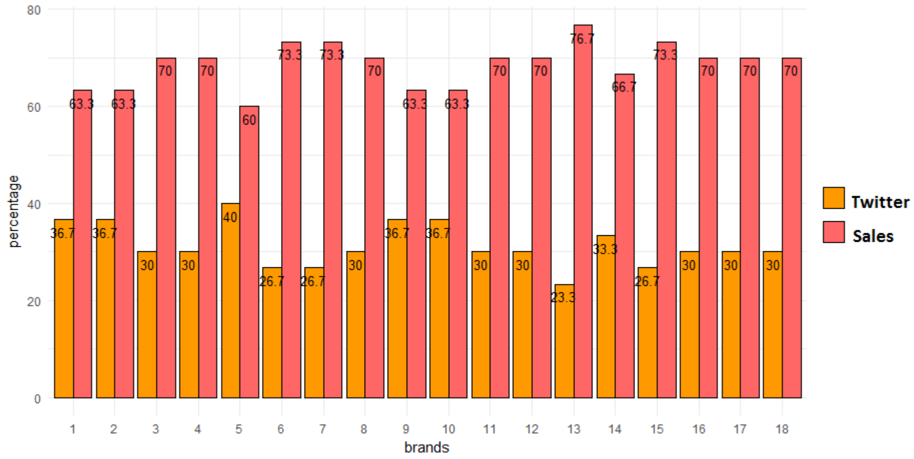


Figure 8.5: Proportions of selected feature types for the feature group S_T.

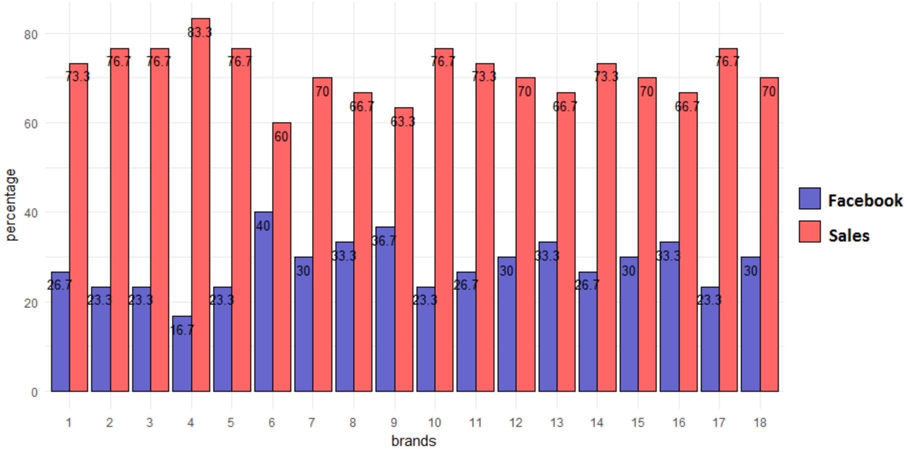


Figure 8.6: Proportions of selected feature types for the feature group S_F.

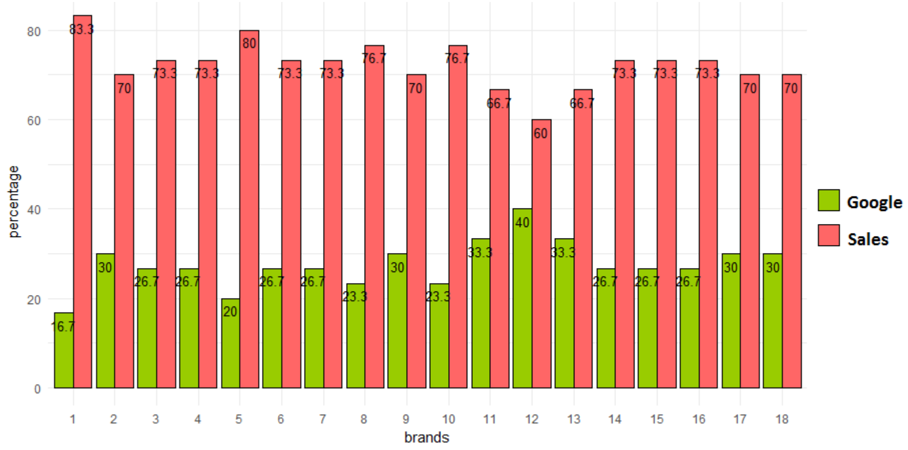


Figure 8.7: Proportions of selected feature types for the feature group S_G.

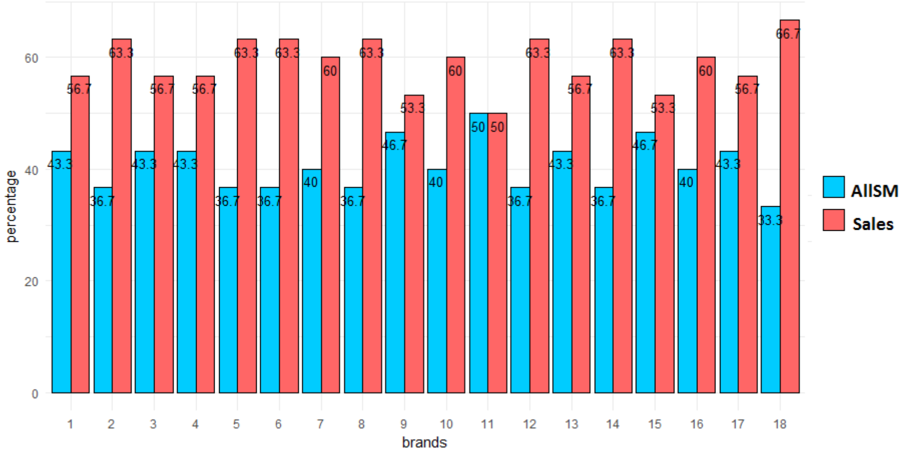


Figure 8.8: Proportions of selected feature types for the feature group S_AllSM.

Figs. 8.9- 8.12 present the contribution that features of each type deliver to the reduction in MSE for each brand.

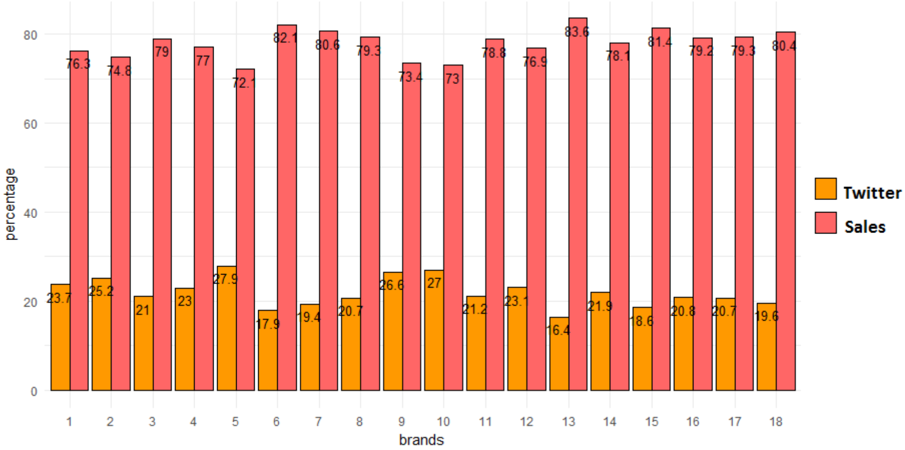


Figure 8.9: Contribution that features of each type deliver to the reduction in MSE for the feature group S_T.

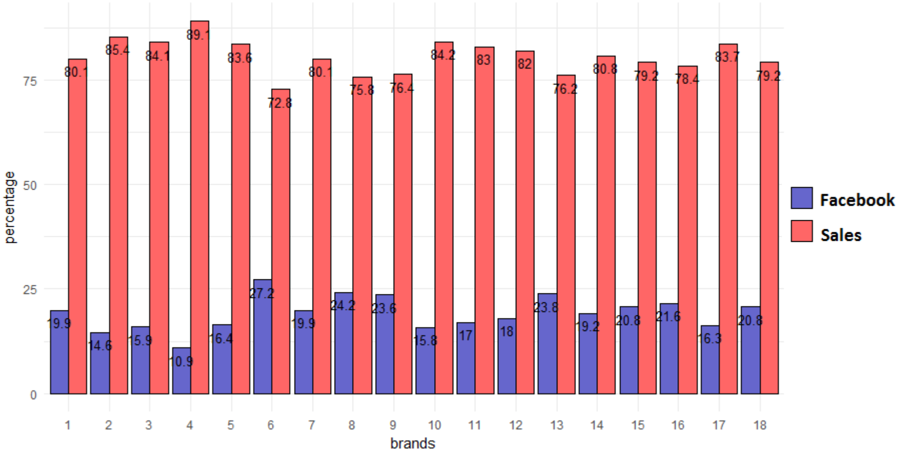


Figure 8.10: Contribution that features of each type deliver to the reduction in MSE for the feature group S_F.

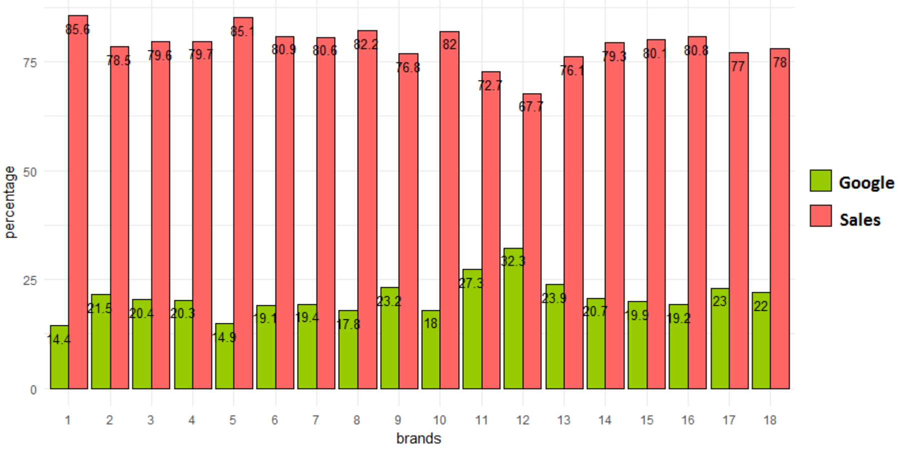


Figure 8.11: Contribution that features of each type deliver to the reduction in MSE for the feature group S_G.

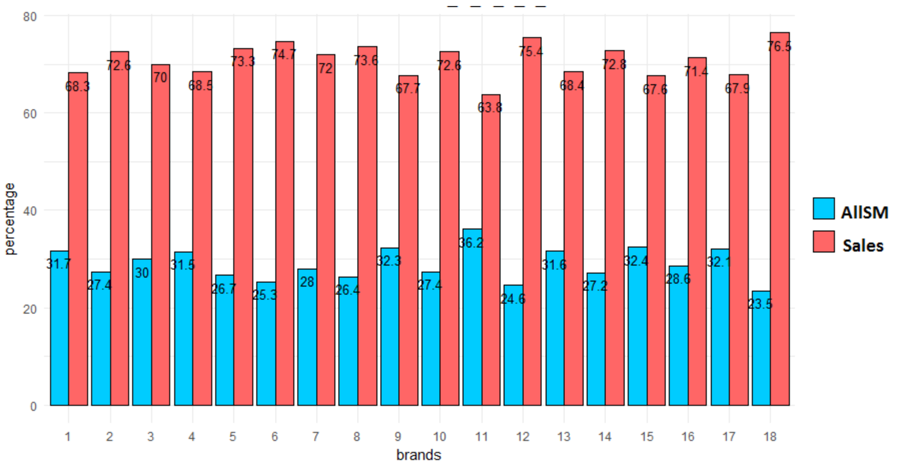


Figure 8.12: Contribution that features of each type deliver to the reduction in MSE for the feature group S_AllSM.

E. Implementation of the Social Media product for Certona

As part of this research work I designed and participated in implementation of a Social Media product for Certona. The purpose and implementation process are described below.

Scope

The purpose of the project was to develop a Social Media product that could be offered to clients as a standalone tool or as a complimentary value proposition to the existing reporting tool. The product should allow to monitor and collect relevant social media information, such as Twitter messages and their sentiment, Facebook posts, Google trends, and present it to clients in the form of signals. Clients should have a possibility to overlay social media signals with other relevant signals, such as revenue, transactions volume, etc. Clients would also be able to view individual messages related to their brand and obtain available information about the people who posted them.

Benefits for Certona

General Outcomes:

- Development of a new revenue stream or an additional value proposition to attract clients;
- Expanding the area of expertise for the business;
- Creating a data asset ;
- Utilising existing expertise (data handling and storage using Hadoop, social media research, reporting tools). The development of the product is therefore:
  - Cost efficient;
  - Time efficient;
  - Requires low effort, but gives high output (added value to clients, possibility to attract more new clients).

Partial Outcomes (during the usage of the product):

- Creating a framework to test clients expectations and needs in social media space;

- Determining the features and direction for future development that have the most impact;
- Qualitative and quantitative feedback of users on product.

Final Outcomes (after 6 months of implementation):

- Impact on attracting new clients;
- Impact on developing new business lines.

Benefits for Clients

- Monitoring relevant social media information related to the brand;
- Monitoring social activity of competitors;
- Monitoring the sentiment related to the brand and possibility to respond quickly;
- Possibility to correlate social media signal with other signals in the report (revenue, transactions, etc.) and create more accurate forecasts;
- Possibility to measure the impact of promotions and marketing campaigns by monitoring social impressions;
- Detection of the most popular tweets/Facebook posts and the most influential users.
- Understanding the demographics of the audience

Implementation in Hadoop

The project consists of the following components (see Fig. 8.13):

1. Twitter Streaming API, as the data source;
2. Apache Flume, for streaming data into HDFS;
3. HDFS (Hadoop Distributed File System), for storage of data;
4. Apache Hive, for transformation and querying of data;
5. Apache Oozie, for automating the data processing.

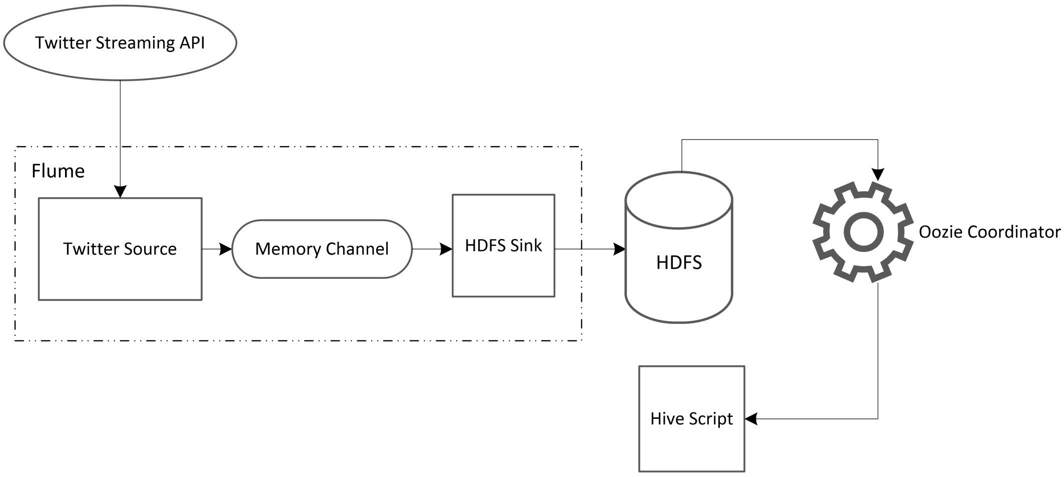


Figure 8.13: Implementation diagram of Twitter streaming.

How it works:

1. Twitter data in JSON format continually streams from the Twitter Streaming API into HDFS via Flume by subscribing to a set of keywords;
2. A Java code invokes Flume to send each tweet to the 'Memory Channel' as an 'event';
3. The Flume channel aggregates and holds these tweets;
4. The Flume sink requests data from the Flume channel at a certain interval;
5. Every interval, a preset number of events gets transferred and placed in an open file in HDFS by the sink;
6. The Flume sink stores the data in HDFS based on the hour of day;
7. An Oozie coordinator job runs hourly, processing the previous hours' JSON data into relational Hive tables for analyst use (see Fig. 8.14 for database design);

Future Work

- Analyse performance metrics to understand clients needs;
- Adjust the value proposition after receiving clients feedback;
- Expand data range to include Facebook, Google Trends, Pinterest, Instagram;
- Generate forecasts based on available data.

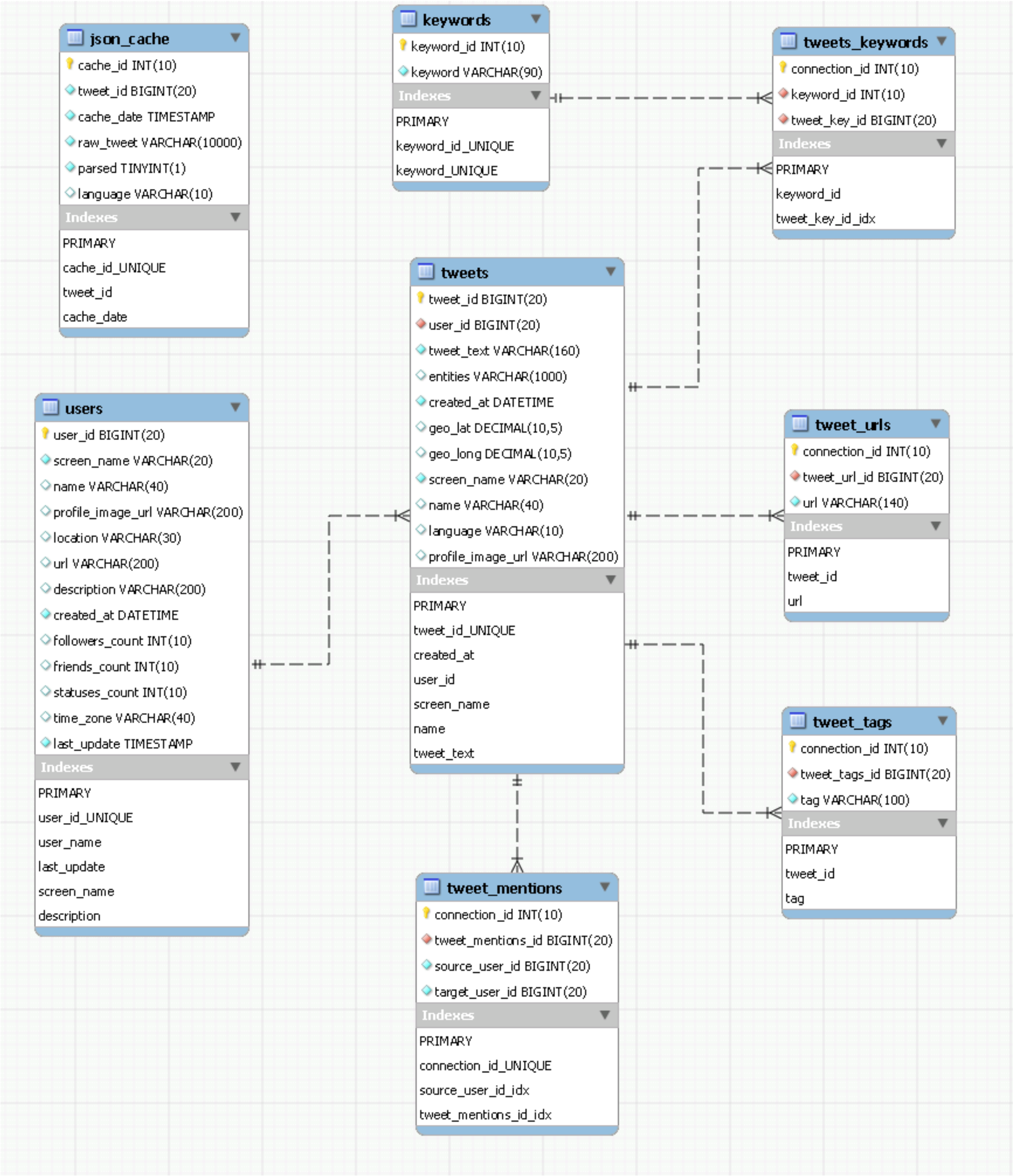


Figure 8.14: Schema for Hive database.

Bibliography

Abbasi, A., Chen, H., and Salem, A. (2008). Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Trans. Inf. Syst.*, 26(3):12:1–12:34.

Abdelhaq, H., Gertz, M., and Armiti, A. (2017). Efficient online extraction of keywords for localized events in twitter. *Geoinformatica*, 21(2):365–388.

Aghabozorgi, S., Shirkhorshidi, A., and Wah, T. (2015). Time-series clustering - a decade review. *Inf. Syst.*, 53(C):16–38.

Aisopos, F., Papadakis, G., Tserpes, K., and Varvarigou, T. (2012). Content vs. context for sentiment analysis: A comparative analysis over microblogs. In *Proceedings of the 23rd ACM Conference on Hypertext and Social Media*, HT '12, pages 187–196, New York, NY, USA. Association for Computing Machinery.

Alanyali, M., Preis, T., and Moat, H. (2016). Tracking protests using geotagged flickr photographs. *PLOS ONE*, 11(3): e0150466.

Allan, J., Carbonell, J., Doddington, G., Yamron, J., and Yang, Y. (1998). Topic detection and tracking pilot study: Final report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pages 194–218, Lansdowne, VA, USA. 007.

Alon, I., Qi, M., and Sadowski, R. J. (2001). Forecasting aggregate retail sales : a comparison of artificial neural networks and traditional methods. *Journal of Retailing and Consumer Services*, 8:147–156.

Alsaedi, N. and Burnap, P. (2015). *Arabic Event Detection in Social Media*, pages 384–401. Springer International Publishing, Cham.

Alsaedi, N., Burnap, P., and Rana, O. (2017). Can we predict a riot? disruptive event detection using twitter. *ACM Transactions on Internet Technology*, 17(2):18:1–18:26.

Ambassador (2013). Social customer service (infographic). `https://www.getambassador.com/blog/social-customer-service-infographic`. Accessed: 06-June-2017.

Amblee, N. and Bui, T. (2007). *The impact of electronic-word-of-mouth on digital microproducts: An empirical investigation of Amazon Shorts*.

Antenucci, D., Cafarella, M., Levenstein, M. C., Ré, C., Shapiro, M. D., Cajner, T., Elsby, M., Nalewaik, J., Tesar, L., West, K., and Wolfers, J. (2014). Using social media to measure labor market flows. `http://www.nber.org/papers/w20010`. Accessed: 2015-04-10.

Appel, O., Chiclana, F., Carter, J., and Fujita, H. (2016). *A Hybrid Approach to Sentiment Analysis with Benchmarking Results*, pages 242–254. Springer International Publishing, Cham.

Aramaki, E., Maskawa, S., and Morita, M. (2011). Twitter catches the flu: detecting influenza epidemics using twitter. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1568–1576.

Asghar, M., Khan, A., Ahmad, S., Qasim, M., and Khan, I. (2017). Lexicon-enhanced sentiment analysis framework using rule-based classification scheme. *PLOS ONE*, 12(2): e0171649.

Asur, S. S. and Huberman, B. A. (2010). Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '10, pages 492–499, Washington, DC, USA. IEEE Computer Society.

Atefeh, F. and Khreich, W. (2015). A survey of techniques for event detection in twitter. *Comput. Intell.*, 31(1):132–164.

Au, K. F., Choi, T. M., and Yu, Y. (2008). Fashion retail forecasting by evolutionary neural networks. *International Journal of Production Economics*, 114(2):615–630.

Bao, T. and Chang, T.-l. S. (2014). Why amazon uses both the new york times best seller list and customer reviews: An empirical study of multiplier effects on product sales from multiple earned media. *Decision Support Systems*, 67:1 – 8.

Bao, T. and Chang, T.-l. S. (2016). The product and timing effects of ewom in viral marketing. *International Journal of Business*, 21(2):99–111.

Barbosa, L. and Feng, J. (2010). Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 36–44, Stroudsburg, PA, USA. Association for Computational Linguistics.

Basili, R., Moschitti, A., and Pazienza, M. T. (2000). Language-Sensitive Text Classification. In *Proceeding of RIAO-00, 6th International Conference "Recherche d'Information Assistee par Ordinateur"*, pages 331–343, Paris, France.

Benamara, F., Irit, S., Cesarano, C., Federico, N., and Reforgiato, D. (2007). Sentiment Analysis: Adjectives and Adverbs are better than Adjectives Alone. In *In Proc of Int Conf on Weblogs and Social Media*.

Benevenuto, F., Magno, G., Rodrigues, T., and Almeida, V. (2010). Detecting spammers on twitter. In *In Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS*.

Berbegal-Mirabent, J., Mas-Machuca, M., and Marimon, F. (2016). Antecedents of online purchasing behaviour in the tourism sector. *Industrial Management & Data Systems*, 116(1):87–102.

Bermingham, A. and Smeaton, A. F. (2011). On using twitter to monitor political sentiment and pre-dict election results. *Proceedings of the Sentiment Analysis where AI meets Psychology Work-shop*.

Berndt, D. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In Fayyad, M. and Uthurusamy, R., editors, *KDD Workshop*, pages 359–370. Association for the Advancement of Artificial Intelligence.

Bessembinder, W., Kahle, K., Maxwell, W., and Xu, D. (2009). New methodology for event studies in bonds. *Review of Financial Studies*, 22:42194258.

Boldt, L. C., Vinayagamoorthy, V., Winder, F., Schnittger, M., Ekran, M., Mukkamala, R. R., Lassen, N. B., Flesch, B., Hussain, A., and Vatrapu, R. (2016). Forecasting nike's sales using facebook data. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 2447–2456.

Box, G. E. P. and Jenkins, G. M. (1976). *Time series analysis: forecasting and control*. Holden-Day series in time series analysis and digital processing. Holden-Day.

Bradlow, E., Gangwar, M., Kopalle, P., and Voleti, S. (2017). The role of big data and predictive analytics in retailing. *Journal of Retailing*, 93(1):79 – 95. The Future of Retailing.

Brown, S. J. and Warner, J. B. (1985). Using daily stock returns: the case of event studies. *Journal of Financial Economics*, 14:3–31.

Bughin, J. (2015). Google searches and twitter mood: nowcasting telecom sales performance. *NETNOMICS: Economic Research and Electronic Networking*, 16(1-2):87–105.

Caropreso, M. F., Matwin, S., and Sebastiani, F. (2001). A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization. In Chin, A. G., editor, *Text Databases and Document Management*, pages 78–102, Hershey, PA, USA. IGI Global.

Carvalho, J., Prado, A., and Plastino, A. (2014). A statistical and evolutionary approach to sentiment analysis. In *Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT) - Volume 02*, WI-IAT '14, pages 110–117, Washington, DC, USA. IEEE Computer Society.

Challis, R. and Kitney, R. I. (1991). Biomedical signal processing (in four parts). part 1 time-domain methods. *Medical and Biological Engineering and Computing*, 28(3):509–524.

Chen, S.-Y., Shin-yi, W., and Jungsun, Y. (2004). *The Impact of Online Recommendations and Consumer Feedback on Sales.*, pages 711–724.

Cheng, T. and Wicks, T. (2014). Event detection using twitter: A spatio-temporal approach. *PLOS ONE*, 9(6):e97807.

Chevalier, J. and Mayzlin, D. (2006). The Effect of Word of Mouth on Sales: Online Book Reviews. *Journal of Marketing Research*, 43(3):345–354.

Choi, H. and Varian, H. (2012). Predicting the present with google trends. *Economic Record*, 88:2–9.

Chong, A., Ch'ng, E., Liu, M., and Li, B. (2017a). Predicting consumer product demands via big data: the roles of online promotional marketing and online reviews. *International Journal of Production Research*, 55(17):5142–5156.

Chong, A., Chng, E., Liu, M., and Li, B. (2017b). Predicting consumer product demands via big data: the roles of online promotional marketing and online reviews. *International Journal of Production Research*, 55(17):5142–5156.

Chong, A., Li, B., Ngai, E., Ch'ng, E., and Lee, F. (2016). Predicting online product sales via online reviews, sentiments, and promotion strategies: A big data architecture and neural network approach. *International Journal of Operations & Production Management*, 36(4):358–383.

Chong, A. and Li, Z. (2014). Demand chain management: Relationships between external antecedents, web-based integration and service innovation performance. *International Journal of Production Economics*, 154(C):48–58.

Chu, C.-W. and Zhang, G. P. (2003). A comparative study of linear and nonlinear models for aggregate retail sales forecasting. *International Journal of Production Economics*, 86(3):217–231.

Clinton, B. (2014). Trendy scraper. `https://github.com/clintonboys/trendy-scraper`. Accessed: 2016-08-15.

Cohen, D. (2015). 50 million small businesses using facebook pages; new communications tools launched. `http://www.adweek.com/digital/50-million-small-businesses-pages-new-communications-tools/`. Accessed: 13-July-2017.

Corley, C. D., Dowling, C., Rose, S., and McKenzie, T. (2013). Social sensor analytics: Measuring phenomenology at scale. In Glass, K., Colbaugh, R., Sanfillippo, A., Kao, A., Gabbay, M., Corley, C., Li, J., Khan, L., Wynne, A., Coote, L., Mao, W., Zeng, D., and Yaghoobi, A., editors, *ISI*, pages 61–66. Institute of Electrical and Electronics Engineers.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. In *Machine Learning*, volume 20, pages 273–297, Hingham, MA, USA. Kluwer Academic Publishers.

Councill, I. G., McDonald, R., and Velikovich, L. (2010). What's great and what's not: Learning to classify the scope of negation for improved sentiment analysis. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, NeSp-NLP '10, pages 51–59, Stroudsburg, PA, USA. Association for Computational Linguistics.

Crane, R. and Sornette, D. (2008). Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences*, 105(41):15649–15653.

Cui, R., Gallino, S., Moreno, A., and Zhang, D. J. (2017). The operational value of social media information. *Production and Operations Management*, pages 313–314.

Culotta, A. (2013). Lightweight methods to estimate influenza rates and alcohol sales volume from twitter messages. *Language resources and evaluation*, 47:1:217–238.

Das, S. and Chen, M. (2001). Yahoo! for amazon: Extracting market sentiment from stock message boards. In *Asia Pacific Finance Association Annual Conf. (APFA)*.

Dash, M. and Liu, H. (1997). Feature selection for classification. In *Intelligent data analysis*, volume 1, pages 131–156. No longer published by Elsevier.

Davidson, R. and MacKinnon, J. G. (2004). *Econometric Theory and Methods*. Oxford University Press.

Davies, L. and Gather, U. (1993). The identification of multiple outliers. *Journal of the American Statistical Association*, 88(423):782–792.

Davis, A. and Khazanchi, D. (2008). An empirical study of online word of mouth as a predictor for multiproduct category ecommerce sales. *Electronic Markets*, 18(2):130–141.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7:1–30.

Derrac, J., Garca, S., Molina, D., and Herrera, F. (2011). A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. *Swarm and Evolutionary Computation*, 1(1):3–18.

Dickey, D. and Fuller, W. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association*, 74(366a):427–431.

Dickinson, B. and Hu, W. (2015). Sentiment analysis of investor opinions on twitter. *Social Networking*, 4:62–71.

Diebold, F. X. and Mariano, R. S. (1995). Comparing Predictive Accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263.

Diederich, J., Kindermann, J., Leopold, E., and Paass, G. (2003). Authorship attribution with support vector machines. In *Applied Intelligence*, volume 19, pages 109–123, Hingham, MA, USA. Kluwer Academic Publishers.

Dijkman, R. M., Ipeirotis, P. G., Aertsen, F., and van Helden, R. (2015). Using twitter to predict sales: A case study. *CoRR*, abs/1503.04599.

Ding, X., Liu, B., and Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 231–240, New York, NY, USA. Association for Computing Machinery.

Dolley, J. (1933). Characteristics and procedure of common stock split-ups. *Harvard Business Review*, (11):316–26.

Dong, X., Mavroeidis, D., Calabrese, F., and Frossard, P. (2015). Multiscale event detection in social media. *Data Mining and Knowledge Discovery*, 29(5):1374–1405.

Du, Q., Fan, W., Qiao, Z., Wang, G., Zhang, X. N., and Zhou, M. (2015). Do facebook activities increase sales? In *AMCIS*. Association for Information Systems.

Dumais, S., Platt, J., Heckerman, D., and Sahami, M. (1998). Inductive learning algorithms and representations for text categorization. In *CIKM '98: Proceedings of the seventh international conference on Information and knowledge management*, pages 148–155, New York, NY, USA. Association for Computing Machinery.

Engle, R. and Granger, C. (1987). Co-integration and error correction: Representation, estimation, and testing. *Econometrica*, 55(2):251–76.

Esuli, A. and Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. `http://www.bibsonomy.org/bibtex/25231975d0967b9b51502fa03d87d106b/mkroell`. Accessed: 2014-07-07.

Fersini, E., Messina, E., and Pozzi, F. (2015). Expressive signals in social media languages to improve polarity detection. In *Information Processing and Management*.

Fersini, E., Messina, E., and Pozzi, F. A. (2014). Sentiment analysis: Bayesian ensemble learning. In *Decision Support Systems*, volume 68, pages 26 – 38.

Flannagan, R. (2014). How important are online customer reviews. `https://nuancedmedia.com/how-important-are-online-customer-reviews/`. Accessed: 18-July-2017.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. In *The Journal of Machine Learning Research*, volume 3, pages 1289–1305. JMLR.org.

Franses, P. H. and Draisma, G. (1997). Recognizing changing seasonal patterns using neural networks. *Journal of Econometrics*, pages 273–280.

Gamboa, A. and Gonalves, H. (2014). Customer loyalty through social networks: Lessons from zara on facebook. *Business Horizons*, 57(6):709–717.

Gao, Y., Zhang, H., Zhao, X., and Yan, S. (2017). Event classification in microblogs via social tracking. *ACM Transactions on Intelligent Systems and Technology*, 8(3):35:1–35:14.

Giles, D. (2013). Ardl models - part ii - bounds tests.

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., and Smith, N. A. (2011). Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 42–47, Stroudsburg, PA, USA. Association for Computational Linguistics.

Go, A., Bhayani, R., and Huang, L. (2009). Twitter sentiment classification using distant supervision. *Processing*, pages 1–6.

Granger, C. and Newbold, P. (1974). Spurious regressions in econometrics. *Journal of Econometrics*, 2(2):111–120.

Granger, C. and Weiss, A. (1983). {TIME} {SERIES} {ANALYSIS} {OF} error-correction {MODELS}. In Karlin, S., Amemiya, T., and Goodman, L., editors, *Studies in Econometrics, Time Series, and Multivariate Statistics*, pages 255 – 278. Academic Press.

Granger, C. W. J. (1981). Some properties of time series data and their use in econometric model specification. *Journal of Econometrics*, 16(1):121–130.

Gruhl, D., Guha, R., Kumar, R., Novak, J., and Tomkins, A. (2005). The predictive power of online chatter. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, KDD '05, pages 78–87, New York, NY, USA. Association for Computing Machinery.

Gur Ali, O. and Pinar, E. (2016). Multi-period-ahead forecasting with residual extrapolation and information sharing - Utilizing a multitude of retail series. *International Journal of Forecasting*, 32(2):502–517.

Guta, M. (March, 2016). Exploring video social network trends. `https:`

//smallbiztrends.com/2016/03/video-social-networks.html. Accessed: 08-08-2017.

Hall, M. (2012). Twitter labelled dataset. https://drive.google.com/file/d/0B1pvkpCwTsiSd1pyTFZkdWVRdEs5Q1NiQW1mRmF1Zw/view. Accessed: 06-Feb-2013.

Hampel, F. R. (1971). A general qualitative definition of robustness. *Ann. Math. Statist.*, 42(6):1887–1896.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *j-J-AM-STAT-ASSOC*, 69(346):383–393.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *Cluster Analysis*, chapter 14.3. New York: Springer-Verlag.

Hatzivassiloglou, V. and McKeown, K. (1997). Predicting the semantic orientation of adjectives. pages 174–181, Madrid, Spain.

Hjalmarsson, E. and Osterholm, P. (2007). Testing for cointegration using the Johansen methodology when variables are near-integrated. International Finance Discussion Papers 915, Board of Governors of the Federal Reserve System (U.S.).

Hogenboom, A., Bal, D., Frasincar, F., Bal, M., de Jong, F., and Kaymak, U. (2013). Exploiting emoticons in sentiment analysis. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, SAC '13, pages 703–710, New York, NY, USA. Association for Computing Machinery.

Hu, M. and Liu, B. (2004). Opinion mining, sentiment analysis, and opinion spam detection. http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html. Accessed: 2014-03-20.

Hu, N., Pavlou, P., and Zhang, J. (2006). *Can online reviews reveal a product's true quality? Empirical findings analytical modeling of online word-of-mouth communication*, volume 2006, pages 324–330.

Hu, W., Wang, H., Qiu, Z., Nie, C., Yan, L., and Du, B. (2017). An event detection method for social networks based on hybrid link prediction and quantum swarm intelligent. *World Wide Web*, 20(4):775–795.

Hu, X., Tang, L., Tang, J., and Liu, H. (2013). Exploiting social relations for sentiment analysis in microblogging. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 537–546, New York, NY, USA. Association for Computing Machinery.

Hudson, S. and Thal, K. (2013). The impact of social media on the consumer decision process: Implications for tourism marketing. *Journal of Travel & Tourism Marketing*, 30(1-2):156–160.

Huhtala, Y., Karkkainen, J., and Toivonen, H. (1999). Mining for similarities in aligned time series using wavelets. volume 3695, pages 150–160.

Hyndman, R. (2016). *Measuring Forecast Accuracy*, pages 1–39. Wiley.

Ikonomakis, M., Kotsiantis, S., and Tampakas, V. (2005). Text classification using machine learning techniques. In *WSEAS Transactions on Computers*, volume 4, pages 966–974.

Imran, M., Castillo, C., Diaz, F., and Vieweg, S. (2015). Processing social media messages in mass emergency: A survey. *ACM Computing Surveys*, 47(4):67:1–67:38.

Imran, M., Castillo, C., Lucas, J., Meier, P., and Vieweg, S. (2014). Aidr: Artificial intelligence for disaster response. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, pages 159–162, New York, NY, USA. Association for Computing Machinery.

Indyk, P., Koudas, N., and Muthukrishnan, S. (2000). Identifying representative trends in massive time series data sets using sketches. In *Proceedings of the 26th International Conference on Very Large Data Bases*, VLDB '00, pages 363–372. Morgan Kaufmann Publishers Inc.

Iqbal, J. and Uddin, M. (2013). Forecasting accuracy of error correction models:

International evidence for monetary aggregate m2. *Journal of International and global Economic Studies*, 6(1):14–32.

Jacobs, P. S. (1992). *Joining Statistics with NLP for Text Categorization.* Accessed: 2014-05-07.

Jansen, D. W. and Wang, Z. (2006). *Evaluating the Fed Model of Stock Price Valuation: An out-of-sample forecasting perspective*, pages 179–204.

Joachims, T. (1999). Transductive inference for text classification using support vector machines. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, pages 200–209, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Johansen, S. (1988). Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control*, 12(2-3):231–254.

Johansen, S. (1991). Estimation and hypothesis testing of cointegration vectors in gaussian vector autoregressive models. *Econometrica*, 59(6):1551–80.

Jurek, A., Mulvenna, M. D., and Bi, Y. (2015). Improved lexicon-based sentiment analysis for social media analytics. *Security Informatics*, 4(1):9.

Kaji, N. and Kitsuregawa, M. (2007). Building lexicon for sentiment analysis from massive collection of html documents. In *EMNLP-CoNLL*, pages 1075–1083. Association for Computational Linguistics.

Kallus, N. (2014). Predicting crowd behavior with big public data. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, pages 625–630, New York, NY, USA. Association for Computing Machinery.

Kalpakis, K., Gada, D., and Puttagunta, V. (2001). Distance measures for effective clustering of arima time-series. In *Proceedings 2001 IEEE International Conference on Data Mining*, pages 273–280.

Keogh, E. and Kasetty, S. (2003). On the need for time series data mining benchmarks: A survey and empirical demonstration. *Data Mining and Knowledge Discovery*, 7(4):349–371.

Keogh, E. and Ratanamahatana, C. A. (2005). Exact indexing of dynamic time warping. *Knowl. Inf. Syst.*, 7(3):358–386.

Keogh, E. and Smyth, P. (1997). A probabilistic approach to fast pattern matching in time series databases. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, KDD'97, pages 24–30. Association for the Advancement of Artificial Intelligence.

Kilian, L. and Ltkepohl, H. (2017). *Vector Error Correction Models*, chapter 3. Cambridge University Press.

Kim, S. and Hovy, E. (2004). Determining the sentiment of opinions. pages 1267–1373, Geneva, Switzerland.

Kim, S. M. and Hovy, E. H. (2007). Crystal: Analyzing predictive opinions on the web. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Com-putational Natural Language Learning*, pages 1056–1064.

Kolchyna, O., Souza, T. T. P., Treleaven, P. C., and Aste, T. (2016). A framework for twitter events detection, differentiation and its application for retail brands. In *2016 Future Technologies Conference (FTC)*, pages 323–331.

Konchitchki, Y. and OLeary, D. E. (2011). Event study methodologies in information systems research. *International Journal of Accounting Information Systems*, 12:99–115.

Kotsakos, D., Trajcevski, G., Gunopulos, D., and Aggarwal, C. (2013). *Time-Series Data Clustering*, chapter 15. Chapman and Hall/CRC.

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. In *Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*, pages 3–24, Amsterdam, The Netherlands, The Netherlands. IOS Press.

Kouloumpis, E., Wilson, T., and Moore, J. (2011). Twitter sentiment analysis: The good the bad and the omg! In Adamic, L. A., Baeza-Yates, R. A., and Counts, S., editors, *ICWSM*. Association for the Advancement of Artificial Intelligence.

Krenker, A., Bester, J., and Kos, A. (2011). *Introduction to the Artificial Neural Networks*. InTech.

Kristoufek, L., Moat, H., and Preis, T. (2016). Estimating suicide occurrence statistics using google trends. *EPJ Data Science*, 5(1):32.

Krumm, J. and Horvitz, E. (2015). Eyewitness: Identifying local events via space-time signals in twitter feeds. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '15, pages 20:1–20:10, New York, NY, USA. Association for Computing Machinery.

Kulkarni, G., Kannan, P. K., and Moe, W. (2012). Using online search data to forecast new product sales. *Decision Support Systems*, 52(3):604–611.

Kwiatkowski, D., Phillips, P., and Schmidt, P. (1991). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? Cowles Foundation Discussion Papers 979, Cowles Foundation for Research in Economics, Yale University.

Ladha, L. and Deepa, T. (2011a). Feature selection methods and algorithms. In *International Journal on Computer Science and Engineering*, volume 3, pages 1787–1797.

Ladha, L. and Deepa, T. (2011b). Feature selection methods and algorithms, international journal on computer science and engineering. In *International Journal on Computer Science and Engineering*, volume 3, pages 1787–1800.

Lassen, N., Madsen, R., and Vatrapu, R. (2014). Predicting iphone sales from iphone tweets. In *2014 IEEE 18th International Enterprise Distributed Object Computing Conference*, pages 81–90.

Lee, K., Lee, B., and Oh, W. (2016). Thumbs Up, Sales Up? The Contingent Effect of Facebook Likes on Sales Performance in Social Commerce. *Journal of Management Information Systems*, 32(4):109–143.

Li, C., Sun, A., and Datta, A. (2012). Twevent: Segment-based event detection from tweets. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 155–164, New York, NY, USA. Association for Computing Machinery.

Li, J., Wen, J., Tai, Z., Zhang, R., and Yu, W. (2016). Bursty event detection from microblog: A distributed and incremental approach. *Concurr. Comput. : Pract. Exper.*, 28(11):3115–3130.

Liao, T. W. (2005). Clustering of time series dataa survey. *Pattern Recognition*, 38:1857–1874.

Ling, C. X. and Sheng, V. S. (2007). Cost-sensitive Learning and the Class Imbalanced Problem. In Sammut, C., editor, *Encyclopedia of Machine Learning*.

Liu, H., Shah, S., and Jiang, W. (2004). On-line outlier detection and data cleaning. *Computers and Chemical Engineering*, 28(9):1635–1647.

Liu, K.-L., Li, W.-J., and Guo, M. (2012). Emoticon smoothed language models for twitter sentiment analysis. In *Proceedings of the National Conference on Artificial Intelligence*, volume 2, pages 1678–1684. cited By 2.

Liu, X., Wang, M., and Huet, B. (2016a). Event analysis in social multimedia: a survey. *Frontiers of Computer Science*, 10(3):433–446.

Liu, Y. (2006). Word of mouth for movies: Its dynamics and impact on box office revenue. *Journal of Marketing*, 70(3):74–89.

Liu, Y., Yu, X., Huang, X., and An, A. (2010). S-plasa+: Adaptive sentiment analysis with application to sales performance prediction. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 873–874, New York, NY, USA. Association for Computing Machinery.

Liu, Z., Huang, Y., and Trampier, J. (2016b). Leds: Local event discovery and summarization from tweets. In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '16, pages 53:1–53:4, New York, NY, USA. Association for Computing Machinery.

Lopez-Novoa, U., Saenz, J., Mendiburu, A., and Miguel-Alonso, J. (2015). An efficient implementation of kernel density estimation for multi-core and many-core architectures. *The International Journal of High Performance Computing Applications*, 29(3):331–347.

Lovallo, D. and Kahneman, D. (2003). Delusions of success: How optimism undersmines executives' decisions. *Harvard Business Review*, 81(7):1–10.

Lovins, J. B. (1968). Development of a stemming algorithm. In *Mechanical Translation and Computational Linguistics 11*, pages 22–31.

Ma, X., Khansa, L., Deng, Y., and Kim, S. (2014). Impact of prior reviews on the subsequent review process in reputation systems. 30:279 – 310.

MacKinlay, A. C. (1997). Event studies in economics and finance. *Journal of Economic Literature*, 35(1):13–39.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In Cam, L. M. L. and Neyman, J., editors, *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press.

Maddala, G. and Kim, I. (1998). *Unit Roots, Cointegration, and Structural Change*. Themes in Modern Econometrics. Cambridge University Press.

Mallat, S. G. (1989). A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, Massachusetts.

Marsland, S. (2011). *Machine Learning: An Algorithmic Perspective*. CRC Press.

Matsubara, Y., Sakurai, Y., Prakash, B. A., Li, L., and Faloutsos, C. (2017). Nonlinear dynamics of information diffusion in social networks. *Association for Computing Machinery Transactions on the Web*, 11(2):11:1–11:40.

Mayzlin, D. (2006). Promotional chat on the internet. *Marketing Science*, 25(2):155–163.

Mentzer, J. and Bienstock, C. (1998). *Sales Forecasting Management: Understanding the Techniques, Systems and Management of the Sales Forecasting Process*. SAGE Publications.

Mitchell, T. M. (1996). *Machine Learning*. McGrwa Hill, New York, New York, NY, USA.

Moat, H., Curme, C., Stanley, H., and Preis, T. (2014). *Anticipating Stock Market Movements with Google and Wikipedia*, pages 47–59. Springer Netherlands, Dordrecht.

Mohammad, S. M., Kiritchenko, S., and Zhu, X. (2013). Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*.

Morinaga, S., Yamanishi, K., Tateishi, K., and Fukushima, T. (2002). Mining product reputations on the web. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 341–349, New York, NY, USA. Association for Computing Machinery.

Moritz, D. (March, 2016). The shift to visual social media 6 tips for business. `http://sociallysorted.com.au/shift-to-visual-social-media-6-tips-for-business-infographic/`. Accessed: 08-08-2017.

Mudinas, A., Zhang, D., and Levene, M. (2012). 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*, WISDOM '12, pages 5:1–5:8, New York, NY, USA. Association for Computing Machinery.

Muhammad, A., Wiratunga, N., and Lothian, R. (2016). Contextual sentiment analysis for social media genres. *Knowledge-Based Systems*, 108:92 – 101. New Avenues in Knowledge Bases for Natural Language Processing.

Nakov, P., Kozareva, Z., Ritter, A., Rosenthal, S., Stoyanov, V., and Wilson, T. (2013). Semeval-2013 task 2: Sentiment analysis in twitter. *Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, 2:312–320.

Narayanan, V., Arora, I., and Bhatia, A. (2013). Fast and accurate sentiment classification using an enhanced naive bayes model. In Yin, H., Tang, K., Gao, Y., Klawonn, F., Lee, M., Weise, T., Li, B., and Yao, X., editors, *Intelligent Data Engineering and Automated Learning IDEAL 2013*, volume 8206 of *Lecture Notes in Computer Science*, pages 194–201. Springer Berlin Heidelberg.

Ngai, E., Moon, K.-l., Lam, S., Chin, E., and Tao, S. (2015). Social media models, technologies, and applications: An academic review and case study. *Industrial Management & Data Systems*, 115(5):769–802.

Nielsen, F. . (2011). A new anew: evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages 718 in CEUR Workshop Proceedings*, pages 93–98.

Niennattrakul, V. and Ratanamahatana, C. (2007). *Inaccuracies of Shape Averaging Method Using Dynamic Time Warping for Time Series Data*, pages 513–520. Springer Berlin Heidelberg, Berlin, Heidelberg.

Norman, G. and Streiner, D. (2008). *Biostatistics: The Bare Essentials*. Pmph USA Ltd Series. B.C. Decker.

Nowotny, T., Rospars, J.-P., Martinez, D., Elbanna, S., and Anton, S. (2013). Machine learning for automatic prediction of the quality of electrophysiological recordings. In *PlOS ONE*, volume 8(12): e80838.

Ohana, B. and Tierney, B. (2009). Sentiment classification of reviews using sentiwordnet. `http://www.bibsonomy.org/bibtex/2443c5ba60fab3ce8bb93a6e74c8cf87d/bsc`.

Pak, A. and Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In Chair), N. C. C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh*

International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta. European Language Resources Association (ELRA).

Pang, B. and Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the ACL*.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 79–86, Stroudsburg, PA, USA. Association for Computational Linguistics.

Parker, J. (2017). *Regression with Nonstationary Variables*, chapter 4, pages 65–66. unpublished manuscript, Reed College.

Paul, M. J. and Dredze, M. (2011). You are what you tweet: Analyzing twitter for public health. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 265–272.

Pesaran, M. H., Shin, Y., and Smith, R. J. (2001). Bounds testing approaches to the analysis of level relationships. *Journal of Applied Econometrics*, 16:289–326.

Petrović, S., Osborne, M., and Lavrenko, V. (2012). Using paraphrases for improving first story detection in news and twitter. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 338–346, Stroudsburg, PA, USA. Association for Computational Linguistics.

Petrovic, S., Osborne, M., McCreadie, R., Macdonald, C., Ounis, I., and Shrimpton, L. (2013). Can twitter replace newswire for breaking news?

Phillips, P. C. B. and Perron, P. (1988). Testing for a unit root in time series regression. *Biometrika*, 75(2):335.

Phuvipadawat, S. and Murata, T. (2010). Breaking news detection and tracking in twitter. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 03*, WI-IAT '10, pages 120–123, Washington, DC, USA. IEEE Computer Society.

Piantadosi, S. T. (2014). Zipfs word frequency law in natural language: A critical review and future directions. In *Psychonomic Bulletin and Review*, volume 21, pages 1112–1130. Springer US.

Pick, T. (2015a). 21 spectacular seo and search marketing stats and facts. Accessed: 23-August-2017.

Pick, T. (2015b). 31 sensational social media marketing and pr stats and facts. `http://webbiquity.com/social-media-marketing/31-sensational-social-media-marketing-and-pr-stats-and-facts/`. Accessed: 23-August-2017.

Piovani, D., Grujic, J., and Jensen, H. (2015). Forecasting systemic transitions in high dimensional stochastic complex systems. 633:012001.

Polanyi, L. and Zaenen, A. (2006). Contextual Valence Shifters. In Croft, W. B., Shanahan, J., Qu, Y., and Wiebe, J., editors, *Computing Attitude and Affect in Text: Theory and Applications*, volume 20 of *The Information Retrieval Series*, pages 1–10. Springer Netherlands.

Poria, S., Cambria, E., and Gelbukh, A. (2015). Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. *Conference on Empirical Methods in Natural Language Processing*.

Porter, M. (2002). Snowball: Quick introduction. `http://snowball.tartarus.org/texts/quickintro.html`. Accessed: 2014-10-16.

Porter, M. F. (1980). An algorithm for suffix stripping. In *Program*, volume 14, pages 130–137.

Pozzi, F. A., Maccagnola, D., Fersini, E., and Messina, E. (2013). Enhance user-level sentiment analysis on microblogs with approval relations. In Baldoni, M., Baroglio, C., Boella, G., and Micalizio, R., editors, *AI*IA*, volume 8249 of *Lecture Notes in Computer Science*, pages 133–144. Springer.

Prangnawarat, N., Hulpus, I., and Hayes, C. (2015). Event analysis in social media using clustering of heterogeneous information networks. In *FLAIRS Conference*.

Preis, T. and Moat, H. (2014). Adaptive nowcasting of influenza outbreaks using google searches. *Royal Society Open Science*, 1(2).

Preis, T. and Moat, H. (2015). *Early Signs of Financial Market Moves Reflected by Google Searches*, pages 85–97. Springer International Publishing, Cham.

Preis, T., Moat, H. S., and Stanley, H. E. (2013). Quantifying trading behavior in financial markets using google trends. *Sci. Rep.*, 3.

Qian, S., Zhang, T., Xu, C., and Hossain, M. (2015). Social event classification via boosted multimodal supervised latent dirichlet allocation. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 11(2):27:1–27:22.

Quinlan, J. R. (1986). Induction of decision trees. *Mach. Learn.*, 1(1):81–106.

Radinsky, K. and Horvitz, E. (2013). Mining the web to predict future events. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 255–264, New York, NY, USA. Association for Computing Machinery.

Rajaraman, A. and Ullman, J. D. (2011). *Mining of Massive Datasets*. Cambridge University Press, New York, NY, USA.

Ramos, P., Santos, N., and Rebelo, R. (2015). Performance of state space and arima models for consumer retail sales forecasting. *Robotics and Computer-Integrated Manufacturing*, 34:151–163.

Ranco, G., Aleksovski, D., Caldarelli, G., and Grcar, M. (2015). The effects of twitter sentiment on stock price returns. *PLOS ONE*, 10(9): e0138441.

Raskutti, B., Ferrá, H. L., and Kowalczyk, A. (2001). Second order features for maximising text classification performance. In *Proceedings of the 12th European Conference on Machine Learning*, EMCL '01, pages 419–430, London, UK, UK. Springer-Verlag.

Ratanamahatana, C. and Keogh, E. (2005). *Three Myths about Dynamic Time Warping Data Mining*, pages 506–510.

Reuter, T. and Cimiano, P. (2012). Event-based classification of social media streams. In *Proceedings of the 2Nd ACM International Conference on Multimedia Retrieval*, ICMR '12, pages 22:1–22:8, New York, NY, USA. Association for Computing Machinery.

Rijsbergen, C. J. V. (1979). *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 2nd edition.

R.J., H. and Koehler, A. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, pages 679–688.

Rocchio, J. J. (1971). Relevance feedback in information retrieval. In Salton, G., editor, *The Smart retrieval system - experiments in automatic document processing*, pages 313–323. Englewood Cliffs, NJ: Prentice-Hall.

Rogers, D. (1992). A review of sales forecasting models most commonly applied in retail site evaluation. 20.

Rosner, B. (1983). Percentage points for a generalized esd many-outlier procedure. *Technometrics*, 25(2):165–172.

Rozenshtein, P., Anagnostopoulos, A., Gionis, A., and Tatti, N. (2014). Event detection in activity networks. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 1176–1185, New York, NY, USA. Association for Computing Machinery.

Saif, H., He, Y., and Alani, H. (2012). Semantic sentiment analysis of twitter. In *Proceedings of the 11th International Conference on The Semantic Web - Volume Part I*, ISWC'12, pages 508–524, Berlin, Heidelberg. Springer-Verlag.

Sakaki, T., Okazaki, M., and Matsuo, Y. (2010). Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 851–860, New York, NY, USA. Association for Computing Machinery.

Sakoe, H. and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49.

Salton, G. and McGill, M. J. (1983). McGraw Hill Book Co.

Sano, Y., Yamada, K., Watanabe, H., Takayasu, H., and Takayasu, M. (2013). Empirical analysis of collective human behavior for extraordinary events in the blogosphere. *Phys. Rev. E*, 87:012805.

Schapire, R. E. and Singer, Y. (2000). BoosTexter: A Boosting-based System for Text Categorization. In *Machine Learning*, volume 39, pages 135–168.

Schivinski, B. and Dabrowski, D. (2016). The effect of social media communication on consumer perceptions of brands. *Journal of Marketing Communications*, 22(2):189–214.

Schneider, M. and Gupta, S. (2016). Forecasting sales of new and existing products using consumer reviews: A random projections approach. *International Journal of Forecasting*, 32(2):243 – 256.

Schulz, A., Schmidt, B., and Strufe, T. (2015). Small-scale incident detection based on microposts. In *Proceedings of the 26th ACM Conference on Hypertext &#38; Social Media*, HT '15, pages 3–12, New York, NY, USA. Association for Computing Machinery.

Sen, A. (2008). The us fashion industry: A supply chain review. *International Journal of Production Economics*, 114(2):571 – 593. Special Section on Logistics Management in Fashion Retail Supply Chains.

Severyn, A. and Moschitti, A. (2015). Twitter sentiment analysis with deep convolutional neural networks. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 959–962, New York, NY, USA. Association for Computing Machinery.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, 27.

Shao, M., Li, J., Chen, F., Huang, H., Zhang, S., and Chen, X. (2017). An efficient approach to event detection and forecasting in dynamic multivariate social media networks. In *Proceedings of the 26th International Conference on World Wide*

Web, WWW '17, pages 1631–1639, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Shcherbakov, M. V., Brebels, A., Shcherbakova, N. L., Tyukov, A. P., T., A., Janovsky, R., and Kamaev, V. A. (2013). A survey of forecast error measures. *World Applied Sciences Journal*, 24.

Shepard, D. (1968). A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings of the 1968 23rd ACM National Conference*, ACM '68, pages 517–524, New York, NY, USA. Association for Computing Machinery.

Shoutly (2014). Consumers trust social media platforms for product recommendations. `https://www.shoutly.com/en/press/consumers-trust-social-media-platforms-product-recommendations`. Accessed: 25-August-2017.

Sornette, D., Deschatres, F., Gilbert, T., and Ageon, Y. (2004). Endogenous versus exogenous shocks in complex networks: An empirical test using book sale rankings. *Physical Review Letters*, 93.

Souza, T. T. P. and Aste, T. (2016). A nonlinear impact: evidences of causal effects of social media on market prices.

Souza, T. T. P., Kolchyna, O., Treleaven, P. C., and Aste, T. (2016). *Twitter Sentiment Analysis Applied to Finance: A Case Study in the Retail Industry*, chapter 23. Mitra, G. and Yu, X.

Sprenger, T. O., Sandner, P. G., Tumasjan, A., and Welpe, I. M. (2014). News or noise? using twitter to identify and understand companyspecific news flow. *Journal of Business Finance and Accounting*, 41:791–830.

St Louis, C. and Zorlu, G. (2012). Can twitter predict disease outbreaks? BMJ `http://www.bmj.com/content/344/bmj.e2353`. Accessed: 2015-10-03.

Statista (2017). Most famous social network sites worldwide as of august 2017, ranked by number of active users (in millions). `https://www.statista.com/statistics/272014/`

global-social-networks-ranked-by-number-of-users/. Accessed: 14-August-2017.

Stojanovski, D., Strezoski, G., Madjarov, G., and Dimitrovski, I. (2015). *Twitter Sentiment Analysis Using Deep Convolutional Neural Network*, pages 726–737. Springer International Publishing, Cham.

Stone, P. J. and Hunt, E. B. (1963). A computer approach to content analysis: Studies using the general inquirer system. In *Proceedings of the May 21-23, 1963, Spring Joint Computer Conference*, AFIPS '63 (Spring), pages 241–256, New York, NY, USA. Association for Computing Machinery.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. In *Comput. Linguist.*, volume 37, pages 267–307, Cambridge, MA, USA. MIT Press.

Tan, C.-M., Wang, Y.-F., and Lee, C.-D. (2002). The use of bigrams to enhance text categorization. In *INF. PROCESS. MANAGE*, pages 529–546.

Thapen, N., Simmie, D., and Hankin, C. (2016). The early bird catches the term: combining twitter and news data for event detection and situational awareness. *Journal of Biomedical Semantics*, 7(1):61.

Thompson, J. E. (1988). More methods that make little difference in event studies. *Journal of Business Finance and Accounting*, 15:77–86.

Tong, R. (2001). An operational system for detecting and tracking opinions in on-line discussions. In *Working Notes of the SIGIR Workshop on Operational Text Classification*, pages 1–6, New Orleans, Louisianna.

Toshniwal, D. and Joshi, R. (2005). Using cumulative weighted slopes for clustering timeseries data. *GESTS Intl Trans. Computer Science and Engr.*, 20.

Tsay, R. (2005). *Analysis of Financial Time Series*. John Wiley & Sons, second edition.

Tukey, J. (1977). Exploratory data analysis. *Addison-Wesley*.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. (2010). Predicting elections with twit-ter: What 140 characters reveal about political sentiment.

Proceedings of the Fourth Interna-tional AAAI Conference on Weblogs and Social Media, pages 178–185.

Turner, P. (2000). *Aggregate advertising, sales volume and relative prices in the long run*, volume 7, pages 505–508.

Turney, P. D. (2002). Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 417–424, Stroudsburg, PA, USA. Association for Computational Linguistics.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY, USA.

Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley, New York, NY, USA.

Ventosa-Santaularia, D. (2009). Spurious regression. *Journal of Probability and Statistics*, 2009(6):1–27.

Vieweg, S., Hughes, A. L., Starbird, K., and Palen, L. (2010). Microblogging during two natural hazards events: What twitter may contribute to situational awareness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1079–1088, New York, NY, USA. Association for Computing Machinery.

Virmani, V. (2004). Unit root tests: Results from some recent tests applied to select indian macroeconomic variables. IIMA Working Papers WP2004-02-04, Indian Institute of Management Ahmedabad, Research and Publication Department.

Wang, C. and Wang, X. (2000). Supporting content-based searches on time series via approximation. In *Proceedings. 12th International Conference on Scientific and Statistica Database Management*, pages 69–81.

Wang, G., Sun, J., Ma, J., Xu, K., and Gu, J. (2014). Sentiment classification: The contribution of ensemble learning. In *Decision Support Systems*, volume 57, pages 77 – 93.

Wang, X., Smith, K., and Hyndman, R. (2006). Characteristic-based clustering for time series data. *Data Min. Knowl. Discov.*, 13(3):335–364.

WeAreSocial (2017). Digital in 2017: global overview. https://wearesocial.com/special-reports/digital-in-2017-global-overview. Accessed: 11-July-2017.

Weiss, S. M., Apte, C., Damerau, F. J., Johnson, D. E., Oles, F. J., Goetz, T., and Hampp, T. (1999). Maximizing text-mining performance. In *IEEE Intelligent Systems*, volume 14, pages 63–69, Piscataway, NJ, USA. IEEE Educational Activities Department.

Welch, C. (June, 2017). Facebook crosses 2 billion monthly users. The Verge. Vox Media https://www.theverge.com/2017/6/27/15880494/facebook-2-billion-monthly-users-announced. Accessed: 27-06-2017.

Wenbin, H., Huan, W., Chao, P., Huanle, L., and Bo, D. (2017). An event detection method for social networks based on link prediction. *Information Ssytems*, 71:16–26.

Weng, J. and Lee, B.-S. (2011). Event detection in twitter.

Wenjing, D., Bin, G., and Whinston, A. (2008). Do online reviews matter? - an empirical investigation of panel data. *Decision Support Systems*, 45(4):1007–1016.

Werner, W. and Murray, Z. F. (2004). Is all that talk just noise? The information content of internet stock message boards. *Journal of Finance*, 59(3):1259–1294.

Wiebe, J. (2000). Learning subjective adjectives from corpora. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, pages 735–740. Association for the Advancement of Artificial Intelligence.

Wiebe, J. and Wilson, T. (2002). Learning to disambiguate potentially subjective expressions. pages 112–118, Taipei, Taiwan.

Wiebe, J., Wilson, T., and Cardie, C. (2005). Annotating expressions of opinions and emotions in language. In *Language Resources and Evaluation*, volume 39, pages 164–210.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *hltemnlp2005*, pages 347–354, Vancouver, Canada.

Wooldridge, J. (2008). *Multiple regression analysis: Estimation*. South-Western College Pub, 4 edition.

Wright, A. (2009). Mining the web for feelings, not facts. `http://www.nytimes.com/2009/08/24/technology/internet/24emotion.html?_r=0`. Accessed: 2014-10-03.

Wu, L. and Brynjolfsso, E. (2015). *The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales*, pages 89–118. University of Chicago Press.

Xi, X., Keogh, E., Shelton, C., Wei, L., and Ratanamahatana, C. (2006). Fast time series classification using numerosity reduction. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 1033–1040, New York, NY, USA. Association for Computing Machinery.

Yang, J. and Leskovec, J. (2011). Patterns of temporal variation in online media. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 177–186, New York, NY, USA. Association for Computing Machinery.

Yang, X., Ghoting, A., Ruan, Y., and Parthasarathy, S. (2012). A framework for summarizing and analyzing twitter feeds. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 370–378, New York, NY, USA. Association for Computing Machinery.

Yang, Z., Li, Q., Liu, W., Ma, Y., and Cheng, M. (2017). Dual graph regularized nmf model for social event detection from flickr data. *World Wide Web*, 20(5):995–1015.

Yin, J., Karimi, S., and Lingad, J. (2014). Pinpointing locational focus in microblogs. In *Proceedings of the 2014 Australasian Document Computing Symposium*, ADCS '14, pages 66:66–66:72, New York, NY, USA. Association for Computing Machinery.

Yin, J., Lampert, A., Cameron, M., Robinson, B., and Power, R. (2015). Using social media to enhance emergency situation awareness. In *Proceedings of the*

24th International Conference on Artificial Intelligence, IJCAI'15, pages 4234–4238. Association for the Advancement of Artificial Intelligence.

You, Q. and Luo, J. (2013). Towards social imagematics: Sentiment analysis in social multimedia. In *Proceedings of the Thirteenth International Workshop on Multimedia Data Mining*, MDMKDD '13, pages 3:1–3:8, New York, NY, USA. Association for Computing Machinery.

Yule, G. U. (1926). Why do we sometimes get nonsense-correlations between time-series? *Journal of the Royal Statistical Society*, 89:1–63.

Zhang, C., Zhou, G., Yuan, Q., Zhuang, H., Zheng, Y., Kaplan, L., Wang, S., and Han, J. (2016). Geoburst: Real-time local event detection in geo-tagged tweet streams. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 513–522, New York, NY, USA. Association for Computing Machinery.

Zhang, D. (2003). Question classification using support vector machines. In *In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 26–32. Association for Computing Machinery.

Zhang, G. P. (2009). Neural networks for retail sales forecasting. In *Encyclopedia of Information Science and Technology*, pages 2806–2810.

Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., and Liu, B. (2011). Combining lexicon-based and learning-based methods for twitter sentiment analysis. In *Technical report*. Hewlett Packard Labs.

Zhao, J., Dong, L., Wu, J., and Xu, K. (2012). Moodlens: an emoticon-based sentiment analysis system for chinese tweets. In *The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, Beijing, China, August 12-16, 2012*, pages 1528–1531.

Zhao, L., F., C., and Lu, C.T. andRamakrishnan, N. (2016). Multi-resolution spatial event forecasting in social media. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 689–698.

Zheng, X., Zeng, Z., Chen, Z., Yu, Y., and Rong, C. (2015). Detecting spammers on social networks. *Neurocomputing*, 159:27 – 34.

Zhou, X. and Chen, L. (2014). Event detection over twitter social media streams. *The VLDB Journal*, 23(3):381–400.

Zhou, Y., Xu, H., and Lei, L. (2016). Event detection based on interactive communication streams in social network. In *Proceedings of the 9th EAI International Conference on Mobile Multimedia Communications*, MobiMedia '16, pages 54–57, ICST, Brussels, Belgium, Belgium. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

Zhu, L., Galstyan, A., Cheng, J., and Lerman, K. (2014). Tripartite graph clustering for dynamic sentiment analysis on social media. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '14, pages 1531–1542, New York, NY, USA. Association for Computing Machinery.

Zhu, X. and Guo, D. (2017). Urban event detection with big data of taxi od trips: A time series decomposition approach. *Transactions in GIS*, 21(3):560–574.

Zhu, Y., Wang, X., Zhong, E., Liu, N., Li, H., and Yang, Q. (2012). Discovering spammers in social networks. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, AAAI'12, pages 171–177. Association for the Advancement of Artificial Intelligence.