



Universidade do Minho
Escola de Engenharia

Nuno Miguel da Rocha Oliveira

Mining Social Media Sentiment to
Forecast Stock Market Behavior

Nuno Miguel da Rocha Oliveira
Mining Social Media Sentiment to
Forecast Stock Market Behavior



Universidade do Minho
Escola de Engenharia

Nuno Miguel da Rocha Oliveira

Mining Social Media Sentiment to
Forecast Stock Market Behavior

Tese de Doutoramento
Tecnologias e Sistemas de Informação

Trabalho efetuado sob a orientação do
Professor Doutor Paulo Alexandre Ribeiro Cortez
Professor Doutor Professor Nelson Manuel de Pinho
Brandão Costa Areal


STATEMENT OF INTEGRITY

I hereby declare having conducted my thesis with integrity. I confirm that I have not used plagiarism or any form of falsification of results in the process of the thesis elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

Universidade do Minho, 31/01/2017

Full name: Nuno Miguel da Rocha Oliveira

Signature: 

ACKNOWLEDGEMENTS

This PhD thesis was a long and difficult journey, composed by many arduous challenges and several laborious attempts to solve them. Resilience and commitment were essential to make this thesis possible. Many people were critical to transform this lengthy process constituted by many errors and frustrations into a continuous process of learning and improvement.

My supervisor, Prof. Paulo Cortez, was a fundamental support since the beginning of this project. The inestimable expertise, rigor and dedication of Prof. Paulo Cortez guided this work in the right way. My willingness to test new challenges was always nourished by the constant ambitious and enthusiastic attitude of Prof. Paulo Cortez. Moreover, his patience, confidence and incessant encouragement kept me always motivated. I consider Prof. Paulo Cortez as a model of the gifted and dedicated supervisor and researcher.

My co-supervisor, Prof. Nelson Areal, provided us with his intelligence and extreme competence. Prof. Nelson Areal is perfectly suited to this project because he complements his recognized expertise in Finance with an impressive computer proficiency. The valuable advices of Prof. Nelson Areal were crucial to raise the quality level of this thesis.

The faculty and colleagues of the Doctoral Program Doctoral Program on Information Systems and Technologies were important for me to acquire the necessary knowledge for the execution of this project.

The ALGORITMI Center and the School of Economics and Management provided the required computational resources to implement the experiments.

The cooperation with StockTwits was important to our project. StockTwits kindly provided us their valuable data, promoted our work and gave us useful feedback.

The diverse anonymous reviewers of our publications made several helpful suggestions that enhanced our dissertation.

Finally, my family was my anchor. I am fortunate to have a devoted family that gives me constant support and love. I dedicate this thesis to my beloved family.

ABSTRACT

Mining Social Media Sentiment to Forecast Stock Market Behavior

This thesis proposes a novel and fast procedure for creating stock market lexicons based on statistical measures applied over a vast set of labeled messages from a stock market microblog (StockTwits). Using StockTwits, we show that the new lexicons are competitive for measuring investor sentiment when compared with six popular lexicons.

This thesis also presents a robust methodology to assess the value of microblogging data to forecast stock market variables: returns, volatility and trading volume of diverse indices and portfolios. The methodology uses sentiment and attention indicators extracted from microblogs. Such indicators were obtained using a large Twitter data set and the proposed financial microblog lexicon. The methodology also includes the usage of survey indices, several forms to aggregate sentiment indicators, a Kalman Filter to merge microblog and survey sources, a realistic rolling windows evaluation, several Machine Learning methods and the Diebold-Mariano test to validate if the sentiment and attention based predictions are valuable when compared with an autoregressive baseline.

Experimental results show that Twitter sentiment and posting volume were relevant for forecasting the returns of the S&P 500 index, portfolios of lower market capitalization and some industries. Additionally, Kalman Filter sentiment was informative for the forecasting of returns. Moreover, Twitter and Kalman Filter sentiment indicators were useful for the prediction of some survey sentiment indicators. These results confirm the utility of microblogging data for financial decision support systems, allowing the prediction of stock market behavior and providing a valuable alternative for existing survey measures with advantages (e.g., fast and cheap creation, daily frequency).

Análise Automática do Sentimento de Redes Sociais para a Previsão do Comportamento dos Mercados Financeiros

Esta tese propõe um novo e rápido procedimento para criar recursos léxicos para mercados financeiros baseado em medidas estatísticas aplicadas num vasto conjunto de mensagens classificadas de um *microblog* para mercados financeiros (*StockTwits*). Utilizando *StockTwits*, demonstrou-se que os novos recursos léxicos são competitivos quando comparados com seis léxicos populares.

Esta tese apresenta ainda uma metodologia robusta para avaliar o valor de dados de *microblogging* para prever variáveis de mercados financeiros: rendibilidades, volatilidade e volume de transação de diversos índices e portefólios. A metodologia usa indicadores de sentimento e atenção extraídos de *microblogs*. Estes indicadores foram obtidos aplicando dados do *Twitter* e o recurso léxico financeiro proposto. A metodologia também inclui o uso de índices de *surveys*, várias formas de agregar os indicadores de sentimento, *Kalman Filter* para combinar dados de *microblogs* e *surveys*, uma avaliação realista de janelas deslizantes, diversos métodos de *Machine Learning* e o teste Diebold-Mariano para validar as previsões em comparação com um modelo auto regressivo.

Os resultados experimentais mostram que o sentimento e o número de mensagens do *Twitter* são relevantes para a previsão das rendibilidades do index S&P 500, portefólios de menor capitalização e algumas indústrias. Adicionalmente, o sentimento extraído pelo *Kalman Filter* foi informativo para a previsão de rendibilidades. Além disso, os indicadores de sentimento do *Twitter* e do *Kalman Filter* foram úteis para prever alguns valores de sentimento de *surveys*. Estes resultados confirmam a utilidade dos dados de *microblogging* para sistemas financeiros de apoio à decisão, permitindo prever o comportamento dos mercados financeiros e fornecendo uma alternativa para medidas de *survey* existentes com vantagens adicionais (e.g., criação rápida e económica).

CONTENTS

Acknowledgements	v
Abstract	vii
Resumo	ix
Contents	xi
List of Figures	xv
List of Tables	xvii
List of Acronyms	xxi
Glossary	xxv
Part I INTRODUCTION AND BACKGROUND	1
1 INTRODUCTION	3
1.1 Motivation	3
1.2 Objectives	6
1.3 Research Methodology	8
1.4 Contributions	14
1.5 Thesis Organization	17
2 BACKGROUND	21
2.1 Introduction	21
2.2 Machine Learning and Data Mining	24
2.2.1 Multiple Regression	26
2.2.2 Neural Networks	29
2.2.3 Support Vector Machines	36
2.2.4 Random Forest	41
2.2.5 Ensembles	44

Contents

2.3	Sentiment Analysis	47
2.3.1	Background	47
2.3.2	Opinion Representation	48
2.3.3	Classification Levels	49
2.3.4	Methodology	50
2.3.5	Features	56
2.3.6	Lexicons	59
2.3.7	Data Sources	63
2.3.8	Applications	65
2.4	Financial Markets: Basic Concepts	67
2.4.1	Background	67
2.4.2	Machine Learning and Data Mining	69
2.4.3	Sentiment and Attention Indicators	73
2.5	Summary	84
Part II	MAIN BODY	89
3	CREATION OF A SPECIALIZED STOCK MARKET LEXICON	91
3.1	Introduction	91
3.2	Automatic Procedure for the Creation of a Stock Market Lexicon	93
3.2.1	Data Pre-Processing	95
3.2.2	Statistical Measures	96
3.2.3	Scores for Affirmative and Negative Contexts	101
3.3	Lexicon Evaluation	101
3.4	Creation of Investor Sentiment Indicators	104
3.5	Experimental Results and Discussion	106
3.5.1	Lexicon Evaluation	106
3.5.2	Correlation with Survey Sentiment Indicators	112
3.6	Conclusions	115

4	THE IMPACT OF MICROBLOGGING DATA FOR STOCK MARKET PREDICTION	119
4.1	Introduction	119
4.2	Microblogging Sentiment and Attention Indicators	125
4.3	Survey Sentiment Indicators	129
4.4	Sentiment Indicators Created by Kalman Filter Procedure	130
4.5	Stock Market Data	131
4.6	Models	135
4.7	Evaluation	140
4.8	Experimental Results and Discussion	142
4.8.1	Prediction of Returns	142
4.8.2	Prediction of Volatility	149
4.8.3	Prediction of Trading Volume	150
4.8.4	Prediction of Survey Sentiment Indicators Using Twitter and KF Sentiment Indicators	151
4.8.5	Summary of the Predictive Results	155
4.9	Conclusions	157
Part III	CONCLUSIONS	163
5	CONCLUSIONS	165
5.1	Overview	165
5.2	Discussion	169
5.3	Future Work	171
	References	173
	Index	201
Part IV	APPENDICES	207
A	CODE	209

Contents

A.1	Short Examples of the Developed R Code	209
B	DATA	215
B.1	Short Example of the Proposed Lexicon	215
B.2	Short Examples of the Analyzed Data	215

LIST OF FIGURES

Figure 1	Thesis Structure	18
Figure 2	Example of linear least squares fitting. Adapted from Hastie et al. (2013)	29
Figure 3	Structure of a feed-forward NN. Adapted from Bishop (2006)	30
Figure 4	Maximum margin hyperplane and support vectors. Adapted from Burges (1998)	37
Figure 5	Utilization of slack variables in ϵ Support Vector Regression. Adapted from Smola and Schölkopf (2004)	40
Figure 6	Decision boundary using an individual decision tree (figure in the left) and an ensemble of decision trees (figure in the right). Adapted from Seni and Elder (2010)	42
Figure 7	SA approaches	55
Figure 8	SA algorithm components	56
Figure 9	SA data sources	65
Figure 10	Domains applying SA	67
Figure 11	Distribution of sentiment data sources	78
Figure 12	Distribution of sentiment analysis approaches	79
Figure 13	Distribution of predicted stock types	81
Figure 14	Distribution of methods applied for financial analysis	82
Figure 15	Results obtained by the literature about the analysis of the predictive value of sentiment and attention for stock market behavior	83
Figure 16	Bullish and bearish word cloud for a stock market lexicon (PMI_{All}).	111

List of Figures

Figure 17	Distribution of sentiment scores of SML for affirmative and negated contexts	113
Figure 18	Schematic of the proposed methodology	124
Figure 19	Sentiment indicator values (KF, AAIL, II, UMCS, Sentix and TWT)	132
Figure 20	Predicted and real values for Enrgy	148
Figure 21	Predicted results for negative values of AAIL (left) and values of II calculated by VA formula using KF indicators (right)	153

LIST OF TABLES

Table 1	Summary of SA literature	54
Table 2	Summary of literature about the analysis of the predictive value of sentiment and attention for stock market behavior	76
Table 3	Intensifiers and Diminishers	100
Table 4	Classification results for the created lexicons with unique context score (in %, best values in bold)	107
Table 5	Paired Student's t-test and the Wilcoxon signed rank test for pairwise comparison of lexicons using unique context scores	108
Table 6	Classification results for the selected lexicon creation method and baseline lexicons (in %, best values in bold)	110
Table 7	Paired Student's t-test and the Wilcoxon signed rank test for pairwise comparison of PMI _{All} method with baseline lexicons	110
Table 8	Examples of lexical terms with different sentiment value in PMI _{All} and SWN	110
Table 9	Classification results for unique and dual context scores (in %, best values in bold)	112
Table 10	Paired Student's t-test and the Wilcoxon signed rank test for pairwise comparison of lexicons using unique and two context scores	113
Table 11	Pearson's correlation values of Twitter sentiment indicators with survey sentiment indicators (\star – p-value < 0.01, \diamond – p-value < 0.05, best correlation values for each survey sentiment value in bold)	114

List of Tables

Table 12	Predictive results for returns of SP500, RSL, DJIA, NDQ, RMRF, SMB, HML and MOM using general sentiment indicators calculated by BullR formula and first difference of posting volume	143
Table 13	Predictive results for returns of SP500, RSL, DJIA, NDQ, RMRF, SMB, HML and MOM using general sentiment indicators (BullR, BI and VA approaches), KF indicators and first difference of the posting volume	144
Table 14	Predictive results for returns of portfolios formed on size using general sentiment indicators (BullR, BI and VA approaches), KF indicators and first difference of the posting volume	146
Table 15	Predictive results for returns of portfolios formed on industries using general sentiment indicators (BullR, BI and VA approaches), KF indicators and first difference of the posting volume	147
Table 16	Predictive results for returns of portfolios formed on industries using sectorial sentiment indicators (BullR, BI and VA approaches) and first difference of the posting volume	149
Table 17	Predictive results for VIX and annualized realized volatility of SP500, RSL, DJIA and NDQ using general sentiment indicators (BullR, BI, VA and AG approaches), KF indicators and first difference of the posting volume	150
Table 18	Predictive results for trading volume of SP500 and DJIA using general sentiment indicators (BullR, BI, VA and AG approaches), KF indicators and first difference of the posting volume	151
Table 19	Prediction of weekly AAI and II values using Twitter sentiment indicators calculated by AA approach	152
Table 20	Prediction of weekly AAI and II values using Twitter sentiment indicators calculated by MA approach	154

Table 21	Prediction of weekly AAI and II values using KF sentiment indicators	155
Table 22	Summary of prediction results (number of models with a p-value < 10% in DMST test for each predicted variable with text based indicators)	156
Table 23	Sample (30 days) of the data frame <i>sent</i> containing sentiment values of daily Twitter indicator, weekly survey indicators (AAI and II) and monthly survey indicators (UMCS and Sentix)	213
Table 24	Short sample (45 entries) of the created microblogging stock market lexicon	216
Table 25	Some tweet information converted from JSON format	218
Table 26	Short sample (10 messages) of StockTwits messages, author supplied label and SA classification	219
Table 27	Short sample (40 trading days) of SP500 data collected from Thomson Reuters Datastream	220
Table 28	Short sample (40 trading days) of PInd returns collected from Prof. Kenneth French webpage	221

ACRONYMS

AA	Aggregate Approach
AAII	American Association of Individual Investors
AG	Agreement measure
AR	Auto-Regressive
BearR	Bearish Ratio
BI	Bullishness Index
BOW	Bag of Words
BullR	Bullish Ratio
DJIA	Dow Jones Industrial Average
DM	Data Mining
DMST	Diebold-Mariano Statistical Test for predictive accuracy
EA	Ensemble Averaging
FIN	Financial Sentiment Dictionaries
GI	Harvard General Inquirer
HML	High Minus Low
IG	Information Gain
II	Investors Intelligence
KF	Kalman Filter
MA	Mean Approach
MAE	Mean Absolute Error

Acronyms

MAPE	Mean Absolute Percentage Error
ME	Maximum Entropy
ML	Machine Learning
MOM	Momentum Factor
MPQA	Multi-Perspective Question Answering
MR	Multiple Linear Regression
MSOL	Macquarie Semantic Orientation Lexicon
NB	Naive Bayes
NDQ	Nasdaq 100
NLP	Natural Language Processing
NMAE	Normalized Mean Absolute Error
NN	Neural Networks
NYSE	New York Stock Exchange
OL	Opinion Lexicon
OLS	Ordinary Least Squares
OOB	Out-of-bag
PInd	Portfolios formed on Industries
PMI	Pointwise Mutual Information
POS	Part of Speech
PSize	Portfolios formed on Size
RF	Random Forest
RMRF	Return of the Market minus the Risk-Free return rate
RSL	Russell 2000
SA	Sentiment Analysis
SIC	Standard Industrial Classification
SMB	Small Minus Big

SML	Stock Market Lexicon
SP500	Standard & Poors 500
SVM	Support Vector Machines
SWN	SentiWordNet
TFIDF	Term Frequency - Inverse Document Frequency
TM	Text Mining
TWT	Twitter sentiment indicators
UMCS	University of Michigan Consumer Sentiment index
URL	Uniform Resource Locator
US	United States
VA	Variation
VAR	Vector Auto-Regression
VIX	Chicago Board Options Exchange Volatility Index

GLOSSARY

autoregressive model	regresses the output variable from a time series on its own lagged values.
cashtag	microblogging term composed by a “\$” character and the respective ticker (e.g., \$AAPL). It is commonly used to refer to a stock.
corpus	large and structured collection of texts.
correlation	measures the degree of relationship between two variables.
Diebold-Mariano test	statistical test to compare the predictive accuracy of two forecasts (Diebold and Mariano, 1995).
holdout split method	divides the dataset into two mutually exclusive subsets: a training set and a test set (Kohavi, 1995).
Kalman filter	system of mathematical equations that obtains an effective recursive computational solution of the least-squares method. It can calculate estimates of past, present and future states even when the exact nature of the modeled system is undetermined (Welch and Bishop, 1995).
market capitalization	total market value of the shares outstanding of a publicly traded company (i.e., share price times the number of shares outstanding).
opinion lexicon	list of words with a sentiment value (e.g., positive, negative).
portfolio	set of financial assets.

Glossary

returns	measure of the relative changes in the security value.
rolling window	scheme that uses a fixed size window of training data that is periodically updated by discarding the oldest data and adding the latest data (Moro et al., 2014).
sentiment analysis	computational analysis of opinion, sentiment and subjectivity in text (Pang and Lee, 2008).
stock	financial asset that represents a share in the ownership of a company, i.e. a claim on a portion of its assets and earnings.
stock market index	measurement of the aggregate value of a set of stocks.
Students t-test	statistical hypothesis test that applies a test statistic following a Student's t-distribution under the null hypothesis. This type of tests can have diverse utilizations such as: assess the difference in mean of two populations, verify whether the mean of a population is the value defined in a null hypothesis, check whether the difference between two measurements has an average value of zero.
trading volume	number of shares traded.
tweet	message sent using Twitter. It is restricted to 140 characters.
volatility	latent measure of total risk associated with a given investment.
Wilcoxon signed rank test	non-parametric statistical hypothesis test used to evaluate the differences of the mean of two dependent samples. It can be applied as an alternative to the dependent t-test when data cannot be assumed to be normally distributed.

Part I

INTRODUCTION AND BACKGROUND

INTRODUCTION

1.1 MOTIVATION

Nowadays, very large volumes of data are supplied from diverse sources (e.g., sensors, mobile phones, social media, online shopping) at different speeds and periodicities (Choi et al., 2016; McAfee et al., 2012; Chen and Zhang, 2014; Chen et al., 2014; Russom, 2011). Some sources deliver data continuously in real time (Russom, 2011) and it is estimated that the volume of business data duplicates every 1.2 years (Chen and Zhang, 2014). Laney (2001) classifies these high volume (large volume), variety (diversity of data types and sources) and velocity (fast generation) data as Big Data.

The analysis of big data permits the extraction of valuable information that may improve decision making and organizational performance (McAfee et al., 2012). The identification of new customers and markets, the creation of new products, a better customer service and more operational efficiency are some potential advantages obtained by big data analytics (Chen and Zhang, 2014).

Social media is an important source of big data. For instance, Facebook has 1.18 billion daily active users¹, there are 1 billion visits monthly to sites with embedded Tweets² and Weibo has 236 million monthly active users³. Users spend a significant part of their time on social media services (Fan and Gordon, 2014). Thus, they disclose

¹ <http://newsroom.fb.com/company-info/> [Accessed on 9 January 2017]

² <https://about.twitter.com/company> [Accessed on 9 January 2017]

³ <http://ir.weibo.com/phoenix.zhtml?c=253076&p=iro1-newsArticle&ID=2145407> [Accessed on 9 January 2017]

substantial information about diverse aspects of their lives in these sites. The analysis of these social media data may allow a deeper understanding of users' attitudes and perceptions such as the assessment of their sentiment (e.g., Bollen et al., 2011), the identification of their interests (e.g., Han and Lee, 2016) or the measurement of users influence (e.g., Tang et al., 2009).

Social media mining can be utilized for various purposes. For instance, it can be used to predict sales (Fan and Gordon, 2014); to manage brand reputation (Gundecha and Liu, 2012; Omand et al., 2012); to create recommendation systems (Gundecha and Liu, 2012); to identify and predict crime (Chen et al., 2014); to improve advertisement by finding influential users to promote their products and by evaluating the reactions to an ad-campaign (Chen et al., 2014; Omand et al., 2012; Gundecha and Liu, 2012); to identify pandemics and to assess mental health (Omand et al., 2012) or to measure users sentiment towards political parties (Omand et al., 2012).

Financial services have been exploring social media data. For instance, investor sentiment and attention indicators can be easily extracted from social media data in order to support stock market decisions. The evaluation of the predictive value of social media data for stock market behavior is an active research topic. Diverse studies have extracted investor sentiment and attention indicators from social media contents and have analyzed their impact on stock market (e.g., Antweiler and Frank, 2004; Das and Chen, 2007; Bollen et al., 2011; Sprenger et al., 2014; Oh and Sheng, 2011).

Indeed, there is a strand of the finance literature (behavioral finance) that argues that sentiment may affect financial prices (Shiller, 2003). Different proxies for investor sentiment have been applied and most of them are not related to social media. Some papers use indirect measures such as economic and financial variables to infer the emotional state of investors. Other papers applied direct sentiment measures derived from surveys of investors' feelings about the stock market. More recently, various studies used investor sentiment and attention indicators automatically extracted from textual contents such as newspapers, message boards or microblogs. The indicators extracted from texts (e.g., Twitter) have many advantages when compared with survey sentiment

indices (e.g., American Association of Individual Investors (AAII), University of Michigan Consumer Sentiment index (UMCS) or Investors Intelligence (II)). The creation of text based sentiment indicators is faster and cheaper, permits higher frequencies (e.g., daily) and may be targeted to a more restrict set of stocks (e.g., stock market indices or individual stocks). Therefore, social media data may permit a more effective creation of predictive models for finance and permit a wider range of experiments for stock market prediction.

Several studies have used social media sentiment to make trading decisions (Schumaker et al., 2012; Chen and Lazer, 2013) or predict useful stock market variables, such as stock prices (Tetlock, 2007; Bollen et al., 2011), price directions (Groß-Klußmann and Hautsch, 2011; Oh and Sheng, 2011), returns (Sprenger et al., 2014; Bollen et al., 2011; Garcia, 2013), volatility (Antweiler and Frank, 2004; Sabherwal et al., 2011) and trading volume (Tetlock, 2007; Antweiler and Frank, 2004). Posting volume on social media services (e.g., microblogs and message boards) has also been applied in the prediction of returns (e.g., Wysocki, 1998; Antweiler and Frank, 2004), trading volume (e.g., Sprenger et al., 2014; Wysocki, 1998) and volatility (e.g., Antweiler and Frank, 2004; Das and Chen, 2007).

Nevertheless, this research topic is recent and has many challenges. For example, the research findings are not consensual. The efficient market hypothesis, proposes that investors act as rational agents and all available information is reflected immediately in stock prices, which implies that assets are traded at their fair value (Fama, 1998). Indeed, there are studies that find scarce evidence of the predictive power of sentiment or attention for stock prices or returns (Tumarkin and Whitelaw, 2001; Antweiler and Frank, 2004; Das et al., 2005; Timmermann, 2008).

Moreover, most forecasting studies that use text extracted sentiment indicators present limitations in terms of lack of a robust evaluation (e.g., no out of sample evaluation, very short test sets). A large majority of the literature applies linear methods (e.g., Multiple Linear Regression (MR), Vector Auto-Regression (VAR)) to predict stock market variables and very few use models more capable of extracting non-linear associations

(e.g., Support Vector Machines (SVM), Neural Networks (NN)). The comparison of these models is very scarce. The dimension of social media data sets has been very low (e.g., less than two years). The usage of statistical tests to evaluate the predictive accuracy is very limited.

The assessment of the predictive power of sentiment on portfolios (e.g., volatility, book to market, size) is frequent on surveys or financial data studies. However, it is inexistent for sentiment extracted from social media and for higher periodicities than weekly. Many studies apply generic lexicons (e.g., Harvard General Inquirer (GI), Multi-Perspective Question Answering (MPQA) or SentiWordNet (SWN)) to extract sentiment indicators. However, these resources may be ineffective for assessing the sentiment of stock market related messages because some words have distinct sentiment value in financial contexts.

The applied investor sentiment indicators (e.g., surveys, financial measures, social media) may contain some noise because they have different characteristics (e.g., sources, frequencies) and usually present distinct values. The aggregation of these different sentiment indicators may produce a more representative investor sentiment value. There are no studies combining social media sentiment with other sources.

The prediction of survey sentiment indices can be important for investors because it may permit an anticipation or be a cheap alternative measure. Nonetheless, there are no studies attempting to forecast these indices.

1.2 OBJECTIVES

The main goal of this thesis is to perform a rigorous assessment of the value of microblogging data for the forecasting of stock market behavior. Microblogging data permit an easy, fast, inexpensive and flexible production of sentiment and attention indicators that can enhance financial decision support systems. For instance, stock market applications may provide real-time personalized sentiment values (e.g., cus-

tomized periodicity and target stocks) while traditional sentiment indicators are much more rigid (e.g., low periodicities and associated to entire countries). However, these procedural advantages may have reduced value if microblogging data have little predictive value for stock market behavior. Therefore, a robust analysis of the informative content of microblogging data for the forecasting of stock market variables is a relevant activity to obtain a precise assessment of the utility of microblogging data for stock market systems.

In this thesis, we pursue the main goal by addressing the following objectives:

- construct a specialized stock market lexicon properly adapted to microblogging stock market contents to allow a rapid and light unsupervised creation of reliable investor sentiment indicators from microblogging data. The few existing financial lexicons are not adapted to microblogging texts.
- execute a rigorous evaluation of the informative content of microblogging data for the forecasting of stock market variables. A robust methodology (e.g., utilization of diverse Machine Learning (ML) models, statistical test of predictive accuracy, large test sets) should be applied in order to obtain solid evaluation results;
- assess the impact of text based sentiment and attention indicators for different stock market variables (e.g., returns, volatility and trading volume) and stock types (e.g., portfolios, aggregated market, indices). For instance, some studies have found that sentiment has higher impact for some portfolios formed on certain characteristics (e.g., size, book to market) but this analysis is very scarce for social media data and high periodicities.
- create a wide range of accurate investor sentiment indicators in order to obtain solid evaluation results and to allow the execution of a large set of experiments. For example, the production of general and sectorial indicators should permit the assessment of the potential different predictive value of these indicators on the respective sectors. Moreover, the combination of diverse sentiment sources may

enable the extraction of a more representative investor sentiment indicator. The inclusion of these diverse indicators on the experiments can allow the identification of the best creation scheme for investor sentiment indicators (e.g., sentiment data sources, microblogging users, aggregation formulas).

- forecast existing survey sentiment indices (AAII and II) using microblogging data. This procedure may enable an acceptable anticipation of those values or an adequate alternative whenever they are unavailable.

1.3 RESEARCH METHODOLOGY

Two paradigms dominate research in the Information Systems discipline (Hevner et al., 2004). The behavioral science paradigm intends to create and justify theories that explain or forecast organizational or human behavior involving the analysis, design, implementation, management and utilization of information systems. The design-science paradigm is a distinct but complementary research approach. It seeks to produce innovations that define the ideas, practices, technical capabilities, and products through which the analysis, design, implementation, management, and utilization of information systems may be effectively and efficiently achieved (Denning, 1997; Tsichritzis, 1997; Hevner et al., 2004). Design-science is essentially a problem solving approach because it creates and evaluates artifacts that must contribute to the solution of heretofore unsolved organizational problems. Such artifacts may be methods (practices and algorithms), models (representations and abstractions), constructs (symbols and vocabulary), and instantiations (implemented systems). While the objective of behavioral science paradigm is the truth provided by information systems theories, the objective of design science paradigm is the utility supplied by information systems artifacts (Hevner and Chatterjee, 2010).

This thesis intends to evaluate the utility of microblogging data for the prediction of stock market behavior. Since the main goal of this research project is to assess the utility

provided by the development and evaluation of information systems artifacts (e.g., predictive models), we consider that design science is the adequate research paradigm.

Design is a sequence of activities that constructs an innovative artifact through several iterations. In each cycle, the artifact is evaluated in order to provide information to enhance the artifact and the research process. This process ends when the artifact permits a satisfactory solution to the identified problem. Hevner et al. (2004) propose a framework of seven guidelines to assist information systems researchers to understand, execute and evaluate design-science research:

Guideline 1: Design as an Artifact

Design process must create an artifact that could be a construct, a model, a method or an instantiation. These artifacts should constitute an innovation to improve the analysis, design, implementation, and use of information systems (Denning, 1997; Tsihritzis, 1997; Hevner et al., 2004). They must be described in a manner that should permit their implementation and utilization by other researchers and practitioners.

The main objective of this project is to evaluate the relevance of microblogging data for the forecasting of stock market behavior. To assess the informative content of microblogging data, several models using diverse microblogging features are created to predict stock market variables and survey sentiment values. For instance, this iterative research process experiments sentiment indicators extracted using various lexical resources (e.g., existing lexicons, novel stock market lexicon), related to different targets (e.g., individual stocks, aggregate market, industries), combining diverse sources (e.g., microblogging data, surveys) and aggregated by different formulas (e.g., bullishness index, bullish ratio). Furthermore, these predictive models are tested for several stock market variables (e.g., returns, volatility, trading volume), stock types (e.g., individual stocks, indices, aggregate market and portfolios) and surveys (e.g., AAIL, II). To ensure a fast and light creation of investor sentiment indicators, this thesis also proposes a novel method to automatically create a specialized stock market lexicon using microblogging data, as detailed in Chapter 3.

Guideline 2: Problem Relevance

Research in information systems aims at acquiring knowledge that allows the development and implementation of technology-based solutions to unsolved and relevant business problems. Design science addresses this objective by producing innovative artifacts. The prediction of stock market behavior is an important domain application area that involves many agents (e.g., investors, companies). This thesis studies the development of alternative forms to forecast stock market variables by using microblogging features in their predictive models. Microblogs are a recent source of investors attention and sentiment indicators because a large community of investors uses these services to share their opinions. The extraction of investor attention and sentiment indicators from microblogging data is more efficient (e.g., cheaper, faster, more flexible) than traditional sources (e.g., surveys). Hence, an adequate application of microblogging data may permit the creation of more effective predictive models for stock market variables. However, the impact of sentiment and attention on stock prices is a controversial topic, as explained in Chapter 2, and thus more research is needed in this area.

The utilization of microblogging data also permits to extend research in this area. For example, microblogging data permits the instant creation of high frequency (e.g., intraday, daily) sentiment indicators for a specific group of stocks. Traditional sources offer a smaller range of research possibilities because they produce low periodicity (e.g., weekly, monthly) indicators usually related to aggregated information for markets or countries. Thus, the higher flexibility of microblogging indicators permits a wider amplitude of possible experiments for the assessment of the impact of sentiment and attention for the forecasting of stock market behavior.

Guideline 3: Design Evaluation

The contributions of the artifacts (e.g., utility, quality, efficacy) should be proved by rigorous evaluation methods. Artifacts may be evaluated in diverse dimensions such as completeness, functionality, accuracy, consistency, reliability, performance, fit with the organization and usability.

This evaluation phase also supplies important feedback to the iterative and incremental design process. The construction of the artifact has many cycles and its continued evaluation provides essential information to perform the necessary improvement activities. Knowledge base contains several evaluation methods that can be used in this phase such as observational (e.g., case study, field study), analytical (e.g., static analysis, dynamic analysis, optimization), experimental (e.g., controlled experiment, simulation), testing (e.g., functional testing, structural testing) and descriptive (e.g., informed argument, scenarios). The evaluation methods should be selected according to the characteristics of the designed artifact.

This thesis applies controlled experiments to assess the utility of microblogging data for stock market variables. We evaluate the forecasting content of these data by comparing the predictive accuracy of models using microblogging features with similar models without microblogging information. Therefore, we consider that microblogging has utility if models using its information are significantly more accurate than baseline models. These predictive improvements are assessed by the Diebold-Mariano Statistical Test for predictive accuracy (DMST) (Diebold and Mariano, 1995).

Guideline 4: Research Contributions

Design-science research must deliver contributions in the areas of the design artifact, design construction knowledge or design evaluation knowledge. The majority of the contributions of design-science research is the designed artifact. This research product should contribute to the solution of existing unsolved problems by extending the knowledge base or applying existing knowledge in a innovative manner. Additionally, the creative creation of novel models, constructs, methods, or instantiations that enlarge and enhance the existing foundations in the design-science knowledge base are valuable contributions. Moreover, the proposal of new evaluation metrics and the innovative development and utilization of evaluation methods (e.g., observational, analytical, experimental) constitute design-science research contributions.

The main research contribution of this dissertation is the utility assessment of microblogging data for stock market behavior prediction. Microblogging data permit a

more rapid, economical and flexible creation of investor attention and sentiment indicators than traditional sources (e.g., surveys). Thus, the utilization of microblogging data may allow the production of more effective predictive models.

Furthermore, this thesis proposes an original method to automatically produce a specialized stock market lexicon properly adapted for Sentiment Analysis (SA) on microblogging stock market conversations. To the best of our knowledge, there were no lexicons adjusted for the creation of investor sentiment indicators from microblogging messages. The application of this lexicon may improve unsupervised SA on microblogging stock market contents. Hence, it can enhance the creation of investor sentiment indicators especially when there is no classified training data and limited computational resources.

Guideline 5: Research Rigor

Research demands an effective utilization of the knowledge base (i.e., research methodologies and theoretical foundations). In the specific case of design-science research, the development and evaluation of the artifact requires the correct application of robust methods in order to produce solid results. Hence, an essential ability of design-science researchers is to select the most adequate techniques to construct the artifact and the most appropriate methods and metrics to evaluate it.

This project uses knowledge from various areas. The assessment of the predictive value of sentiment for stock market behavior is a research topic that already has a considerable dimension. The analysis of this literature was important to understand the rationale for the potential impact of sentiment, to study the applied predictive models and to analyze the respective evaluation methods. However, we intended to include some novel contributions that demanded further study of related areas. For instance, the creation of reliable investor sentiment indicators required a deeper analysis of SA and Text Mining (TM) areas. A better understanding of financial markets allowed the proposal of original hypothesis. Moreover, the application of more robust evaluation methods demanded further knowledge about statistics. Therefore, this extensive analy-

sis was crucial to perform an informed selection of the most proper methods to develop and evaluate the designed artifacts.

Guideline 6: Design as a Search Process

Problem solving can be described as using available means to satisfy desired ends while complying with laws existing in the environment (Simon, 1996). Means are the available group of actions and resources to develop a solution, ends correspond to the objectives and constraints on the solution and laws are incontrollable forces in the environment. However, it is often impossible to determine the means, ends or laws (Vessey and Glass, 1998). When it is possible, the problem is frequently computationally intractable. Therefore, heuristic search strategies are common approaches to produce good and feasible designs that can be applied in the business context.

The design activity involves the creation, usage, and evaluation of heuristic search strategies. It is an iterative search process to find an efficient solution to unsolved problems. The objective is to construct an artifact that contributes to a satisfactory solution without having to specify all possible solutions.

The application of heuristics requires the evaluation of the goodness of the solution. This assessment may be performed by diverse approaches depending on the problem representation. For instance, the utility of the artifact may be demonstrated by obtaining evaluation values close to an optimal solution or by outperforming state of the art solutions.

This thesis is based on a thorough analysis of the literature about the predictive value of sentiment and attention for stock market behavior and a comprehensive review of the state of art about SA methods. This analysis permitted the identification of new hypothesis and inspired the creation of novel models to be tested during this project. Since there is a vast number of possible combinations of features, an important research task was to identify the potentially most informative features that could be iteratively tested. The literature review also allowed the definition of appropriate baseline models and evaluation procedures to be applied in each design cycle. Thus, the designed artifacts were iteratively developed by evaluating the accuracy of both created and

baseline models and by identifying and performing the necessary modifications in each design cycle. This iterative process permitted an effective development of predictive models using microblogging data for stock market behavior and survey indices. A robust evaluation of these models demonstrated that microblogging features can be useful for the forecasting of some stock market variables and survey values. This dissertation also proposed a novel procedure to automatically create a specialized stock market lexicon. The produced lexicon have proved to be more informative for SA on microblogging stock market messages than existing lexicons.

Guideline 7: Communication of Research

Design-science research should be communicated effectively both to technology and management oriented audiences. Technology-oriented audiences require enough detail to allow the implementation and utilization of the described artifact within an adequate organizational environment. Management-oriented audiences demand sufficient detail to decide if the organization should invest in the construction or purchase of the artifact.

This project presented its research work in two JCR Q₁ journals⁴, three international conferences and one book chapter. These publications are described in Section 1.4. The proposed microblog financial lexion was also made publicly available at https://github.com/nunomroliveira/stock_market_lexicon.

1.4 CONTRIBUTIONS

The main contributions of this PhD dissertation are:

- a novel procedure to automatically create a specialized stock market lexicon properly adjusted to microblogging stock market contents. There are very few financial lexicons and the existing ones are not adapted to microblogging texts. Therefore, the produced lexicon should permit a fast and light unsupervised extraction

⁴ First Quartile (Q₁) according to the impact factor of Journal Citation Report (JCR) published by Thomson Reuters.

- of more solid investor sentiment indicators from microblogging data. The proposed procedure uses three adapted statistical measures (e.g., Pointwise Mutual Information (PMI)) and two new complementary statistics on labeled microblogging messages to calculate a stock market sentiment score. To address negation more effectively, sentiment values are also computed for affirmative and negated contexts;
- a rigorous evaluation of the relevance of microblogging data for the prediction of stock market variables by applying a more robust methodology than used in the research literature about this topic. The methodology utilizes a realistic rolling window evaluation, the DMST test for predictive accuracy and the comparison of five ML models: MR, NN, SVM, Random Forest (RF) and Ensemble Averaging (EA);
 - a large number of experiments to assess the potential different impact of sentiment and attention on distinct stock market variables and types. Hence, we tested the prediction of diverse stock market variables (returns, volatility and trading volume) of indices (e.g., Standard & Poors 500 (SP500) and Nasdaq 100 (NDQ), aggregated market and portfolios formed based on industries and size. We note that within the state of the art works, the prediction of daily values of portfolios using social media data is very scarce;
 - new microblogging sentiment indicators based on the newly created specialized microblogging stock market lexicon. These indicators were extracted from a very large Twitter data set, containing approximately 31 million tweets from December 2012 to October 2015 about all stocks traded in United States (US) markets (about 3,800 stocks). The applied data set is larger than most studies using social media data. Moreover, we experimented the production of sentiment indicators for different targets (e.g., aggregated market, industries) and aggregated by diverse formulas (e.g., bullish ratio, variation, agreement);

- an unique daily sentiment indicator aggregating measures of different periodicities and sources: daily microblogging indicators, weekly and monthly survey indices (e.g., AAIL, II, UMCS, Sentix). This latent indicator extracted by a Kalman Filter (KF) procedure may represent a less noisy measure of investor sentiment;
- prediction of existing survey sentiment indices (AAIL and II) using microblogging data, which may allow a satisfactory anticipation of those values or a decent alternative whenever they are unavailable.

This thesis may support further advances in financial applications. For example, real time financial decision support systems can be improved by providing customized sentiment and prediction indicators about selected stocks (e.g., individual stocks, indices) or surveys (e.g., AAIL, II) for different periodicities (e.g., intraday, daily, weekly). Hence, it may contribute to the creation of more effective and flexible financial applications.

The contributions of this PhD dissertation are included in the following publications:

- Journal papers:
 - Nuno Oliveira, Paulo Cortez, and Nelson Areal. The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices. *Expert Systems with Applications*, 73:125–144, 2017 (JCR 2015 Q1 in "Computer Science, Artificial Intelligence").
 - Nuno Oliveira, Paulo Cortez, and Nelson Areal. Stock market sentiment lexicon acquisition using microblogging data and statistical measures. *Decision Support Systems*, 85:62–73, 2016 (JCR 2015 Q1 in "Computer Science, Artificial Intelligence" and Q1 in "Computer Science, Information Systems").
- Conference papers:
 - Nuno Oliveira, Paulo Cortez, and Nelson Areal. Automatic Creation of Stock Market Lexicons for Sentiment Analysis Using StockTwits Data. In *Proceed-*

ings of the 18th International Database Engineering & Applications Symposium, pages 115–123. ACM Press, 2014.

- Nuno Oliveira, Paulo Cortez, and Nelson Areal. On the predictability of stock market behavior using stocktwits sentiment and posting volume. In *Progress in Artificial Intelligence*, volume 8154 of *Lecture Notes in Computer Science*, pages 355–365. Springer Berlin Heidelberg, 2013.
- Nuno Oliveira, Paulo Cortez, and Nelson Areal. Some experiments on modeling stock market behavior using investor sentiment analysis and posting volume from Twitter. In *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics*, article no. 31. ACM Press, 2013
- Book chapters:
 - Nuno Oliveira, Paulo Cortez, and Nelson Areal. Sentiment Analysis of Stock Market Behavior from Twitter Using the R Tool. In *Text Mining and Visualization: Case Studies Using Open-Source Tools*, pages 223–240, CRC Press, 2015.

1.5 THESIS ORGANIZATION

The organization of this thesis is illustrated in Figure 1.

This dissertation consists of four main parts, which include five chapters and 3 appendices. The first part is related with introductory material and includes Chapters 1 (Introduction) and 2 (Background). Chapter 2 describes the main knowledge necessary to define and develop this research project. Section 2.2 explains the relevance of Data Mining (DM) and ML for this topic and details five ML methods adopted in this work. The creation of text based investor sentiment indicators requires robust SA models. Thus, a comprehensive analysis of the state of the art about SA is included in Section 2.3. It contains aspects such as the main SA methodologies, features, data sources, applications and procedures to create lexicons. Section 2.4 presents the basic

Chapter 1. Introduction

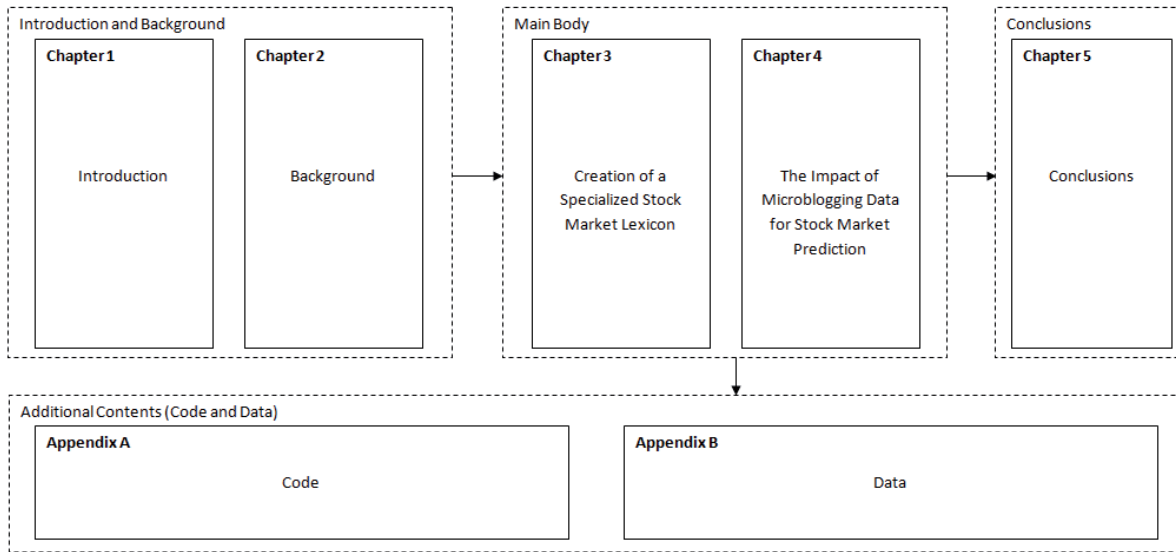


Figure 1.: Thesis Structure

concepts of financial markets, the most common financial applications applying DM and a rigorous literature review about the evaluation of the predictive value of sentiment and attention for stock market behavior. This review supports the identification of the main research opportunities for this thesis.

Chapters 3 and 4 describe the main experimental work (i.e., Main Body part) of this thesis and are based on two journal publications (Oliveira et al., 2016, 2017) made during his PhD project. A novel and fast approach to automatically create stock market lexicons is presented in Chapter 3. This procedure is based on statistical measures applied over a vast set of labeled messages from a specialized stock market microblog, StockTwits. Three adaptations of statistical measures (e.g., PMI), two new complementary statistics and the usage of sentiment scores for affirmative and negated contexts are proposed. The informative content of the created stock market lexicon for SA is evaluated by comparing SA results on StockTwits using this lexicon and six popular lexicons. Moreover, this chapter calculates the correlation values of Twitter investor sentiment indicators with survey sentiment indices.

Chapter 4 proposes a robust methodology to assess the value of microblogging data to forecast stock market variables: returns, volatility and trading volume of diverse

indices and portfolios. The methodology uses sentiment and attention indicators extracted from microblogs (a large Twitter data set is adopted) and survey indices (AAII, II, UMCS and Sentix), diverse forms to daily aggregate these indicators, usage of a KF to merge microblog and survey sources, a realistic rolling windows evaluation, several ML methods and the DMST test to validate if the sentiment and attention based predictions are valuable when compared with an autoregressive baseline. A similar methodology was also applied in the forecasting of two popular survey sentiment indices: AAI, II.

The third part, corresponding to Chapter 5, presents the main conclusions of this thesis, summarizing and discussing the achieved results. Also, it suggests future lines of research. The fourth and final part presents the thesis appendices, which includes short examples of the thesis results, namely the developed R code (Appendix A), proposed lexicon (Appendix B) and analyzed data (Appendix B).

As a final thesis organization note, we highlight that the conference and journal article publication record (see Section 1.4) reflects the true research path that was executed in this PhD work. The first two conference papers (2013) are related with initial attempts to assess the value of microblogging data to forecast stock market behavior. These attempts allowed to identify a weakness related with the quality of the sentiment analysis methods used by related works. This weakness was explored as an opportunity that led to the conference paper published in 2014, where a new lexicon creation method was proposed. Such work was then extended, leading to the first journal article published in 2016. Finally, we enhanced and extended the initial approach (executed in 2013), by adopting the new microblog financial lexicon and many other novel features, leading to the second journal article, accepted in late 2016 and to appear in 2017. In order to reduce some overlap and to present a more coherent writing, in this thesis we opted to include only the journal article related research that appears in the Main Body part (Chapters 3 and 4).

BACKGROUND

2.1 INTRODUCTION

Mining social media data to forecast stock market behavior is a multidisciplinary research topic that can benefit from the knowledge of distinct areas such as Finance, DM, ML, TM, Natural Language Processing (NLP), SA and Statistics.

The analysis of stock market behavior requires some knowledge about financial markets. The transaction of securities and commodities is coordinated by these markets. For instance, companies obtain capital and investors manage their finances using these operations.

Behavioral finance is a strand of finance that argues that sentiment may affect financial prices. According to this research framework, sentiment and attention can be informative for stock market prediction. Moreover, the impact of sentiment and attention may vary on different stock types (e.g., small and young stocks), variables (e.g., returns, volatility) and periodicities (e.g., intraday, daily, monthly). However, this view is controversial and needs further research.

DM provides tools that can be important to find informative patterns from large financial and social media data. For example, they can be used to predict categorical (e.g., price directions) or continuous values (e.g., returns) based on various types of input data such as previous stock market performance (e.g., previous returns), company fundamental data (e.g., cash flow) and indicators aggregating social media activity

(e.g., microblogging sentiment and attention indicators). ML algorithms are particularly useful for this purpose because they are designed to be able to learn from data. A good understanding of the characteristics of existing ML methods (e.g., SVM, NN or RF) permits a better selection and parameterization of efficient methods for each task and may impact the results of this study.

Social media contents have some attributes that distinguish them from most data sources and that impose additional challenges. The computational analysis of text is very difficult because text is unstructured, sparse and high dimensional. There are large amounts of important information that are expressed in text and their human analysis is unfeasible. Therefore, it is relevant to perform an adequate automatic extraction of meaningful information from amorphous text documents. TM is the variant of DM that addresses this type of problems and that is used to perform tasks such as text classification, document summarization, text clustering or entity extraction. Despite the apparent similarity, DM and TM have significant differences (Witten et al., 2011). For instance, data pre-processing and representation of TM models differs from typical DM models. Most TM models use Bag of Words (BOW) representations by characterizing each document by the frequency or presence of its words. However, BOW may ignore some important information for text analysis such as word order and syntactical relations. Thus, other models apply more complex representations by using richer semantic information such as named entities, position information or syntactical associations. NLP tools can be very useful for text representation by performing tasks such as Part of Speech (POS) tagging, entity and relations extraction or disambiguation. Common DM techniques (e.g., NN, SVM, Naive Bayes (NB)) can take these text representations as inputs and extract informative patterns. TM is a burgeoning area due to its novelty and complexity (Witten et al., 2011).

SA is a specific TM task that is particularly important for this research topic. The computational analysis of sentiment in text permits a fast assessment of the investor sentiment. Indeed, the sentiment of large volumes of social media messages related to stock market issues can be automatically extracted using SA models. Moreover, this

assessment may be targeted for various stocks, indices or portfolios by selecting the respective social media messages. Hence, the utilization of SA models permits a rapid and inexpensive creation of personalized sentiment indicators.

The exponential spread of Social Media data has reinforced the utility of TM and SA. Additionally, there are several arguments recommending the utilization of social media data for financial prediction. For instance, microblogging services (e.g., Twitter, Stock-Twits) have large investor communities sharing information about their stock market activities. These users react readily to events and post frequently during trading days. Thus, the analysis of investors social networks interactions can provide robust investor sentiment indicators in a faster, cheaper and more flexible manner than traditional sources (e.g., large-scale surveys). For example, the application of SA in social media data permits an almost instant creation of sentiment indicators according to different investor needs such as various periodicities (e.g., intraday, daily) or stocks (e.g., individual stocks, indices). However, social media text imposes additional challenges for TM. For instance, social media contents often contains specific terminology and poor vocabulary that may weaken the efficiency of standard NLP tools. Also, it often lacks contextual information because many social media messages are short replies to other messages.

The usage of statistics is very important for diverse tasks of this research project. For instance, the application of statistical measures permits the assessment of the informative content of several lexical items for SA (e.g., Information Gain (IG), PMI). Moreover, the selection of the most adequate SA model can be based on various evaluation metrics (e.g., F1 score, accuracy) and statistical tests (e.g., Students t-test and the non-parametric Wilcoxon signed rank test). Furthermore, the predictive value of sentiment and attention indicators for stock market behavior can be evaluated by a pair-wise statistical test of predictive accuracy between models using these features and a baseline model without these features.

Chapter 2. Background

The study of the mentioned areas is important to gather the main scientific and technological background knowledge needed for the definition and execution of this multidisciplinary work. In this chapter, we focused on three main areas:

- DM and ML methods that are essential to create predictive models for stock market variables and important for the production of sentiment indicators.
- SA that is crucial to create robust investor sentiment values by properly processing the peculiarities of textual opinionated contents.
- the basic concepts about financial markets. In this section, we identify some important DM applications in finance and perform an extensive analysis of the literature about the evaluation of the informative value of sentiment and attention for diverse stock market variables. This information is fundamental to define and test hypothesis related to the predictive content of social media data for stock market behavior.

This chapter is organized as follows. Section 2.2 describes the importance of DM and ML for this project and presents a complete characterization of the used DM methods. Section 2.3 presents a comprehensive review about SA, including its main applied methodologies, features, data sources and applications. Section 2.4 exposes the basic concepts of financial markets, an analysis about the main applications of DM in finance and a detailed literature review about the predictive value of sentiment and attention for stock market behavior. This chapter ends with a summary in Section 2.5.

2.2 MACHINE LEARNING AND DATA MINING

The computerization of our activities, the massive utilization of social media and the rapid evolution of data collection and storage technologies has caused an explosive increase of available data. Indeed, high data volume is created everyday in diverse

domains such as business (e.g., sales transactions, customer feedback), telecommunications (e.g., data traffic), engineering (e.g., system performance, engineering observations) and health industry (e.g., medical records, patient monitoring) (Han et al., 2011). The intensive and widespread utilization of search engines and social media has also contributed to the massive creation of data (Han et al., 2011). These huge quantities of data convey useful information that may constitute competitive advantages to organizations. However, the human processing of these large quantities of data is impracticable without the application of tools to automatically extract valuable knowledge from enormous and multiple data sets. DM is the process of discovering valuable patterns from large amounts of raw data (Witten et al., 2011; Turban et al., 2008; Han et al., 2011). These patterns should be advantageous to the organization and have economic value (Witten et al., 2011). DM may represent a strategic tool by offering better conditions for decision making. For instance, it is often applied to decrease fraudulent behavior, identify lucrative customers, discover buying patterns or determine trading strategies (Turban et al., 2008). Hence, the importance of DM has grown considerably.

There are diverse categories of DM algorithms, such as:

- Classification algorithms that predict the correct categorical label of unclassified records based on the analysis of a training data set.
- Regression algorithms are similar to classification methods but are applied to forecast continuous values while classification methods assign class labels.
- Clustering algorithms divide a database into groups of members holding similar characteristics (Turban et al., 2008). These clusters are created by maximizing the similarity between members of each group and by minimizing the similarity between elements of different groups (Han et al., 2011). Unlike regression and classification, clustering does not need labeled training data.
- Mining associations identify relationships between items within data. For instance, it is common to analyze sales transactions in order to discover items that

are sold together (Turban et al., 2008; Han et al., 2011). This analysis can also be temporal in order to extract associations over time.

ML is a relevant field of study for DM. Many techniques used in DM are from this domain. ML studies the development of algorithms capable of learning from data. These algorithms are able to automatically improve their performance based on data in tasks such as the identification of complex patterns and the proposal of effective decisions (Han et al., 2011). This section presents four different regression methods used in this PhD project.

2.2.1 *Multiple Regression*

Linear regression is a simple and important statistical method for numeric forecasting. It has been widely applied in diverse disciplines (e.g., finance, economics) for decades. Linear regression provides accurate predictions for some problems (e.g., linear relationship between target and input variables) and an interpretable description of the relationship between variables (Hastie et al., 2013). Furthermore, linear methods can be used in nonlinear problems by transforming their inputs. Linear regression models express the numeric output as a linear combination of the inputs as:

$$f(X) = \beta_0 + \sum_{j=1}^p X_j \beta_j \quad (1)$$

where β_0 and β_j are the parameters that have to be determined from data and X_j are the independent variables. Therefore, this model is linear in the parameters β and independent variables X . These independent variables may correspond to the raw values of the quantitative inputs but also to transformed inputs (e.g., log or square of inputs, interaction between variables). In this latter situation, the relationship between the target values and the original inputs may be nonlinear.

The target values can be modeled as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad (2)$$

where \mathbf{y} is the N -vector of the target values, \mathbf{X} is a $N \times (p + 1)$ matrix composed by all input vectors (the first position of each input vector is 1), $\boldsymbol{\beta}$ is the $p+1$ parameter vector and the $\boldsymbol{\epsilon}$ vector contains the errors for each instance. Thus, the observed values \mathbf{y} are constituted by a deterministic component ($\mathbf{X}\boldsymbol{\beta}$) and a random part ($\boldsymbol{\epsilon}$) that captures the other factors that influence the observed value. Models containing more than one independent variable are also known as MR models.

The parameters β_j can be interpreted as the partial derivatives of the dependent variable with respect to the corresponding independent variable X_j . Thus, β_j is the expected change in the dependent variable y for a one-unit change in the independent variable X_j holding the other independent variables fixed. These are the values that have to be found. Diverse procedures have been created to estimate these parameters. These methods differ in some aspects such as efficiency of the algorithms, creation of a closed-form solution or the necessary theoretical assumptions to validate the required statistical properties.

The most common estimation method is the Ordinary Least Squares (OLS) method. In this approach, the parameters are determined by minimizing the residual sum of squares (RSS). The residual sum of squares can be written as:

$$RSS(\boldsymbol{\beta}) = \sum_{i=1}^N (y_i - f(x_i))^2 \quad (3)$$

In matrix notation, the function $RSS(\boldsymbol{\beta})$ can be formulated as:

$$RSS(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \quad (4)$$

Differentiating with respect to β produces the normal equations:

$$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) = 0 \tag{5}$$

The minimum always exists because $RSS(\beta)$ is a quadratic function of the β parameters. However, the solution may not be unique (Hastie et al., 2013). If X has full column rank, the unique solution is:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{6}$$

where $\hat{\beta}$ are the fitted β coefficients. Therefore, the predicted values are calculated as:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{7}$$

where $\hat{\mathbf{y}}$ are the predicted values. When matrix \mathbf{X} does not have full column rank, $\mathbf{X}^T\mathbf{X}$ is singular. In these situations, the coefficients $\hat{\beta}$ are not unique. The elimination or recoding of redundant columns are forms to obtain the non-unique representation (Hastie et al., 2013). Figure 2 presents a graphical example of the least-squares fitting.

The quality of the estimates produced by OLS depends on some conditions. OLS estimators are unbiased if errors have conditional mean zero (Greene, 2000). This means that the expected values of the parameters are equal to their true values. The Gauss-Markov theorem defends that OLS is the best linear unbiased estimator for linear regression models in which the errors are uncorrelated, homoskedastic and the expected value is zero (Greene, 2000). When the disturbances (ϵ) are normally distributed, OLS is considered the best unbiased estimator, outperforming non-linear estimators. The presence of linear dependent regressors implies that OLS estimators do not have a unique solution and have infinite variance.

Other estimation techniques may be preferable when data present some characteristics such as heteroscedasticity (e.g. weighted least squares, heteroscedasticity-consistent standard errors), correlated errors (e.g. generalized least squares) or multicollinearity (e.g., ridge regression, lasso regression).

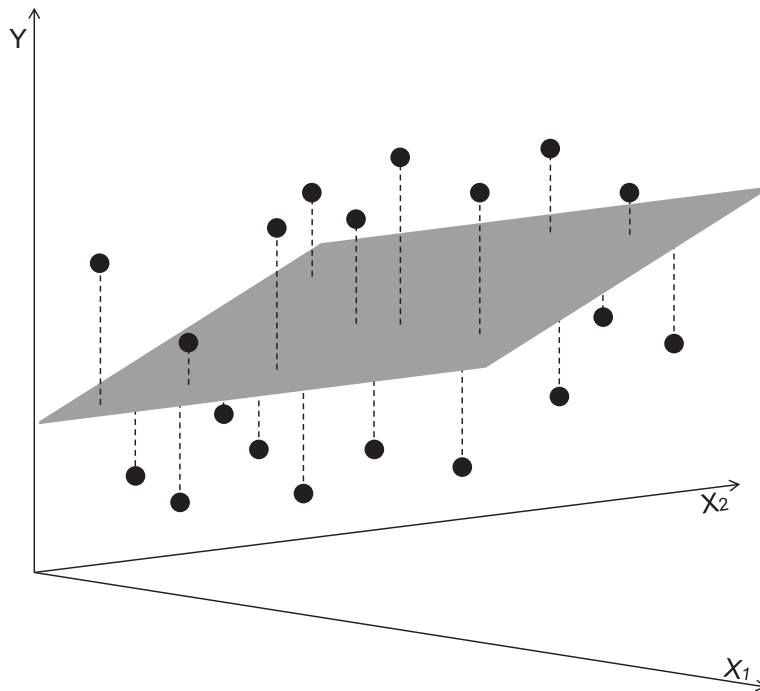


Figure 2.: Example of linear least squares fitting. Adapted from Hastie et al. (2013)

In summary, MR are very fast to learn, less prone to overfitting, easy to interpret and does not need the tuning of hyperparameters. It is extensively applied in financial studies. However, simple linear regression models have limited capacity to approximate nonlinear decision boundaries.

2.2.2 Neural Networks

NN were originated from diverse studies that tried to represent the information processing of biological systems (e.g., McCulloch and Pitts, 1943; Rosenblatt, 1958; Rumelhart et al., 1985). They are a network of connected computational units, organized in layers, with weights associated to these links. During learning, the weights are adjusted in order to improve the prediction of the class labels or the continuous-valued outputs.

Chapter 2. Background

The most popular model is the feed-forward NN. An example of a feed-forward NN is shown in Figure 3.

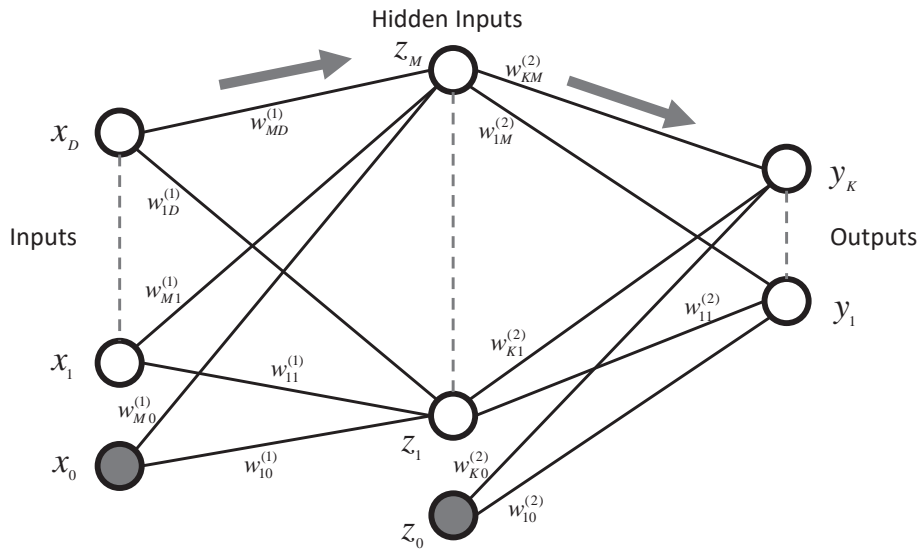


Figure 3.: Structure of a feed-forward NN. Adapted from Bishop (2006)

Feed-forward networks are composed by three types of layers of units. The input layer receives the input values while the output layer provides the predictions produced by the whole network. The intermediate layers are called hidden layers. They take inputs from the previous layer and feed the following layer. Usually there is only one hidden layer, but the number of hidden layers is arbitrary. The nodes of Figure 3 represent the input, hidden and output variables while the weight parameters are illustrated by the links between these nodes. The dark grey nodes x_0 and z_0 are called bias parameters and they do not take any input from the previous layer. The arrows indicate the forward flow of information. The weights of a feed-forward network do not cycle back to a previous layer and its outputs are only based on the current in-

put instance unlike other NN types (e.g., recurrent NN). Each hidden and output unit receives a weighted sum of the outputs of the connected nodes as:

$$a_j = \sum_i w_{ji} z_i \quad (8)$$

where z_i is the output of the unit i that is linked to unit j , and w_{ji} is the weight of the corresponding connection. In this formula, the outputs of biases (i.e., z_0) are fixed at +1 and the outputs of units of the input layer (i.e., z_i used in the calculation of a_j of the first hidden layer) correspond to the input vector of the instance. Then, the sum represented in Formula 8 is converted by an activation function $h(\cdot)$ in the form:

$$z_j = h(a_j) \quad (9)$$

Hidden units of the adopted multilayer perceptron feedforward network generally apply sigmoidal functions (e.g., logistic sigmoid, tanh). These functions map their possibly large input onto a limited interval (e.g., from 0 to 1). For output units, the selection of the activation function is determined by the distribution of target variables and the nature of the data (Bishop, 2006). The logistic sigmoid function is usually used for binary classification problems while the softmax activation function is applied for multiclass problems. The activation function of regression networks is typically the identity so that $y_k = a_k$. The values z_j calculated for the output layer are the outputs of the entire network while the other z_j (i.e., calculated for hidden layers) are the values transmitted to the following layer. The information propagates forward by successively applying Equations 8 and 9, starting on the first hidden layer and finishing in the output layer. This process is called forward propagation. Therefore, the neural network model produces the output values by performing a series of combinations and transformations to a set of input variables. The linear combinations are controlled by adjustable weight parameters that can be learned during training. A NN may classify more than two classes by creating an output unit for each class.

Although NN are also known as multilayer perceptron, they apply differentiable activation functions while the perceptron uses the step-function with discontinuous nonlinearities. This characteristic is very important because it allows the training of the NN parameters through the backpropagation algorithm.

NN are considered universal approximators. For instance, feed-forward networks with only one hidden layer and sigmoidal activation functions can closely approximate any continuous function on a compact input domain (Cybenko, 1989). However, the proper parameter values have to be found. Since an analytical solution to the minimization of the error function is impracticable, NN problems are generally solved by iterative numerical procedures (Bishop, 2006).

The backpropagation algorithm iteratively learns the weights of the network by minimizing an error function. The modifications of the weights are propagated in the backwards direction, starting in the output layer. In each iteration, it is necessary to evaluate the derivatives of the error function with respect to each weight and apply them in the calculation of the weights adjustments. Rumelhart et al. (1985) proposed a simple and effective scheme for applying the gradient descent method for this purpose. An efficient form to evaluate these derivatives is to propagate the errors backwards through the network as we will describe later. Then, the activation values of the network are recalculated by applying the forward propagation using the new weight values. This process is repeated until the termination conditions are satisfied.

The network structure has to be defined before training. For instance, the user must decide the number of units in each layer and the number of hidden layers. The dimensionality of the data set usually determines the number of input and output units (Bishop, 2006). Also, a single hidden layer is frequently enough (Witten et al., 2011). Therefore, the number of units of the hidden layer is commonly the parameter of the network structure that has to be found. The appropriate number of hidden units is usually determined by experimentation.

The normalization of the input values may accelerate the learning phase. Additionally, discrete-valued attributes should be represented by one input unit per domain

value. For instance, an attribute having four possible values should be represented by four distinct input units (Han et al., 2011). Moreover, the weights should be initialized to small random numbers.

The choice of the error function depends on the output unit activation function (Bishop, 2006). Therefore, the sum-of-squares error function is related to linear outputs that are usually used for regression. The cross-entropy error function is associated to logistic sigmoid outputs that are generally applied for binary classification. The multi-class cross-entropy error function is related to softmax outputs that are commonly used for multiclass classification. However, the error function frequently has many local minimums and the backpropagation algorithm may not determine the global minimum. It may be important to compute and compare diverse local minima to discover a good solution.

The derivative of the error function with respect to each weight can be formulated as:

$$\frac{\partial E_n}{\partial w_{ji}} = \delta_j z_i \quad (10)$$

where E_n is the error of the current input instance, w_{ji} corresponds to the weight associated to the connection from node i to node j , z_i is the output of node i and δ_j designates the error of the unit j . Since z_i values were already calculated during the forward propagation phase, we only need to compute δ_j for each unit of the network. For the output units, the derivative is calculated identically for the sum-of-squares, cross-entropy or the multiclass cross-entropy error functions. The formula is:

$$\delta_k = y_k - t_k \quad (11)$$

where y_k is the value produced by output unit k and t_k is the respective target value. The δ values for the hidden units are calculated as:

$$\delta_j = h'(a_j) \sum_k w_{kj} \delta_k \quad (12)$$

Chapter 2. Background

where $h'(a_j)$ is the derivative of the activation function taking a_j as input. Therefore, the calculation of δ values for the hidden units is facilitated by propagating the δ values from units of higher layers. The backpropagation algorithm can also evaluate other derivatives such as the Jacobian and Hessian matrices.

The batch learning requires the processing of the entire training set before updating the weights. In each step, the algorithm sums the derivatives calculated for all training instances as:

$$\frac{\partial E}{\partial w_{ji}} = \sum_n \frac{\partial E_n}{\partial w_{ji}} \quad (13)$$

Then, the weights are modified using these accumulated values. This method is known as gradient descent or steepest descent. However, there are more efficient batch methods (e.g., conjugate gradients) that ensure that the error function always diminishes at each step unless it is at a minimum (Bishop, 2006).

The adjustment of the connection weights can be performed by different forms. The simplest approach is in the form:

$$\mathbf{w}(\tau + 1) = \mathbf{w}(\tau) - \eta \nabla E(\mathbf{w}(\tau)) \quad (14)$$

where \mathbf{w} is a vector of weights, τ is the iteration number and η is the learning rate. The weight update follows the direction of the negative gradient and the magnitude of this modification is proportional to the value of the learning rate and the derivative. The learning rate is useful to escape from local minimum and to discover the global minimum. The value of the learning rate influences the training process. The convergence will be too slow, if the learning rate is too small. However, the search may oscillate between unsatisfactory solutions if the learning rate is too large (Han et al., 2011; Witten et al., 2011).

Most training methods evaluate model performance using the same training set. Therefore, the error metrics of overfitted models may continue decreasing with respect to the training data set when they are already increasing with respect to independent

data. Indeed, it is likely to make a model diverge substantially in these evaluations by training it for a very long time. The early stopping approach uses a holdout set to evaluate model performance during training to address overfitting. Therefore, early stopping ends training at the lowest error value with respect to the holdout set, in order to obtain a good generalization performance (Bishop, 2006). There are other termination conditions for the backpropagation algorithm such as (Han et al., 2011):

- the last alterations to the weights are smaller than a chosen threshold;
- the error metrics of the last training iteration are inferior to a selected threshold;
- the algorithm has already performed a predetermined number of steps.

NN have their advantages and disadvantages. For instance, NN have high tolerance to noisy data and are well adjusted for continuous values (Han et al., 2011). They are more flexible and able to extract nonlinear associations than MR models. Moreover, NN algorithms permit the utilization of parallelization techniques, which accelerates the computation process (Han et al., 2011). In many situations, the obtained model is substantially more compact than SVM while having similar generalization performance (Bishop, 2006).

However, the training time of NN is long, which makes its application infeasible to some problems (Han et al., 2011). Error functions of NN usually have many local minimums and they often do not determine the global minimum unlike other methods such as SVM (Witten et al., 2011; Bishop, 2006). They can have diverse hyperparameters to be tuned and are more prone to overfitting than other models such as MR. NN have in general an unsatisfactory interpretability despite some techniques that extract rules from NN models (Han et al., 2011; Witten et al., 2011).

2.2.3 Support Vector Machines

SVM is a class of learning algorithms originated from the work of Vapnik and his collaborators (e.g., Vapnik and Lerner, 1963; Vapnik and Chervonenkis, 1964, 1974). These algorithms apply linear models that can obtain nonlinear boundaries. Input data may be transformed into a higher dimensional space using nonlinear mapping. Since a straight line in this new space does not appear straight in the original space, a nonlinear boundary may be represented in the original space by creating a linear model in the new space (Witten et al., 2011).

For classification purposes, the linear model tries to find the hyperplane that best separates the data set and that best classifies new instances. Usually, the larger the separation between two classes, the lower the generalization error. Therefore, the algorithm seeks the hyperplane that has the largest margin between classes. Figure 4 presents an example of the maximum-margin hyperplane and respective support vectors.

The maximum-margin hyperplane is exemplified by the line between H1 and H2 hyperplanes in Figure 4. The support vectors are the closest training instances to the maximum-margin hyperplane and they lie on H1 and H2 hyperplanes. The decision boundary is stable because there are few support vectors and the other training vectors can be discarded without modifying the linear model. Since overfitting is motivated by decision boundaries that are too flexible, this algorithm is less prone to overfitting. Linearly separable data sets can be found by:

$$\begin{aligned} & \text{minimize } \frac{1}{2} \|\mathbf{w}\|^2 \\ & \text{subject to } \mathbf{x}_i \cdot \mathbf{w} + b \geq +1 \text{ for } y_i = +1 \\ & \qquad \qquad \mathbf{x}_i \cdot \mathbf{w} + b \leq -1 \text{ for } y_i = -1 \end{aligned}$$

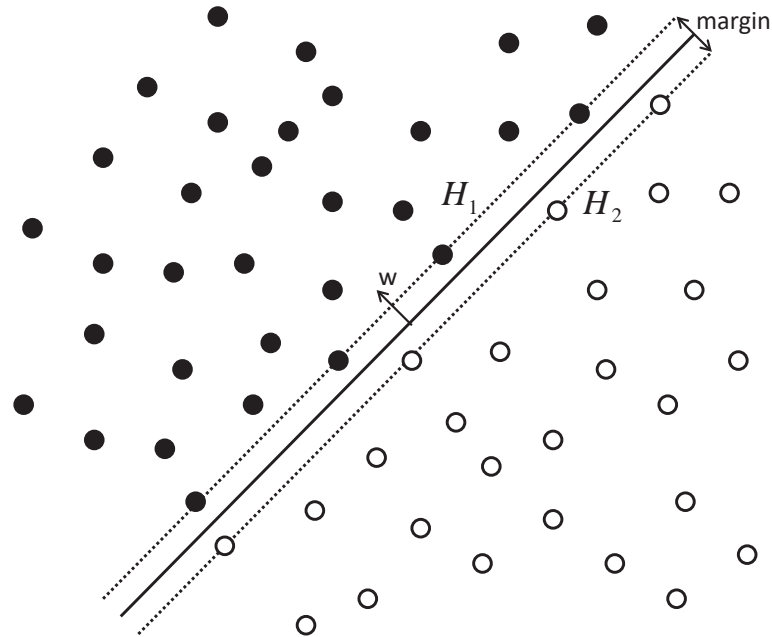


Figure 4.: Maximum margin hyperplane and support vectors. Adapted from Burges (1998)

where \mathbf{x}_i are the training instances, \mathbf{w} is the normal vector to the hyperplane and $y_i \in \{-1, 1\}$ is the respective class value. The constraints can be aggregated as:

$$y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 \geq 0 \quad \forall i \quad (15)$$

The training points for which the equality holds in Equation 15 are the support vectors. The shortest distance between training vectors and the separating hyperplane is $\frac{1}{\|\mathbf{w}\|}$ and the margin is $\frac{2}{\|\mathbf{w}\|}$ (Burges, 1998). The Lagrangian formulation of the mentioned minimization problem permits the identification of the maximum-margin hyperplane and the classification of new instances as:

$$\text{sgn}\left(b + \sum_{i \text{ is a support vector}} \alpha_i y_i \mathbf{x}_i \cdot \mathbf{a}\right) \quad (16)$$

where y_i is the class value of the support vector \mathbf{x}_i , \mathbf{a} corresponds to a test instance, and b and α_i are the parameters that were calculated by the algorithm. This equation only needs the training points that are support vectors. The algorithm computed α_i values superior to zero for support vectors and α_i equal to zero for the remaining training points. Each test instance is classified according to the obtained sign in Equation 16.

In the non-separable case, the constraints are modified to allow a separation with some errors. The soft margin classification is formulated as (Cortes and Vapnik, 1995):

$$\begin{aligned} & \text{minimize} && \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_i \tilde{\zeta}_i \\ & \text{subject to} && y_i(\mathbf{x}_i \cdot \mathbf{w} + b) \geq 1 - \tilde{\zeta}_i \\ & && \tilde{\zeta}_i \geq 0 \quad \forall i \end{aligned}$$

where $\tilde{\zeta}_i$ are positive slack variables that permit the existence of some training points outside the respective area. The parameter C is selected by the user and defines the trade-off between the influence of training errors and the flatness of the function. C is also an upper bound of α_i .

The nonlinear cases are addressed by transforming the original input data into a new high dimensional feature space. Then, the problem is solved just like in the linear setting. However, the nonlinear transformation may increase dramatically the computational complexity because it can create a large number of coefficients. The classification and training tasks imply dot products involving one multiplication and one addition for each attribute. Therefore, these operations in a high-dimensional space are very demanding. However, the application of the kernel trick (Boser et al., 1992) decreases significantly the computational complexity. The application of a kernel function to implicitly map input data into high-dimensional feature spaces is mathematically equivalent and avoids various operations. A function $K(\mathbf{X}, \mathbf{Y})$ is a kernel function if it can be written as $K(\mathbf{X}, \mathbf{Y}) = \Phi(\mathbf{X}) \cdot \Phi(\mathbf{Y})$, where the function Φ can map an instance into a high-dimensional space. The utilization of kernel functions avoids a large number of

mathematical operations by performing the dot products in a lower dimensional space (Witten et al., 2011; Han et al., 2011). Three possible kernel functions are (Han et al., 2011):

- Polynomial kernel of degree h : $K(\mathbf{X}, \mathbf{Y}) = (\mathbf{X} \cdot \mathbf{Y} + 1)^h$
- Gaussian radial basis function kernel: $K(\mathbf{X}, \mathbf{Y}) = \exp(-\|\mathbf{X} - \mathbf{Y}\|^2 / 2\sigma^2)$
- Sigmoid kernel: $K(\mathbf{X}, \mathbf{Y}) = \tanh(\kappa \mathbf{X} \cdot \mathbf{Y} - \delta)$

The decision function for nonlinear transformations changes from Equation 16 to:

$$\text{sgn}(b + \sum_{i \text{ is a support vector}} \alpha_i y_i K(\mathbf{x}_i, \mathbf{a})) \quad (17)$$

For regression, the algorithm needs some modifications. The ϵ -Support Vector Regression algorithm (Vapnik, 1995) tolerate deviations less than ϵ . The objective is to compute the flattest function able to obtain errors inferior to ϵ for all training points. This problem can be written as (Smola and Schölkopf, 2004):

$$\begin{aligned} & \text{minimize} \quad \frac{1}{2} \|\mathbf{w}\|^2 \\ & \text{subject to} \quad y_i - \mathbf{x}_i \cdot \mathbf{w} - b \leq \epsilon \\ & \quad \quad \quad \mathbf{x}_i \cdot \mathbf{w} + b - y_i \leq \epsilon \end{aligned}$$

When it is necessary to allow deviations greater than ϵ , the algorithm may include the slack variables already described in the non-separable data situation. Therefore, these slack variables capture everything above ϵ while errors smaller than ϵ are discarded. Figure 5 illustrates this situation. Only points outside the ϵ area increase the

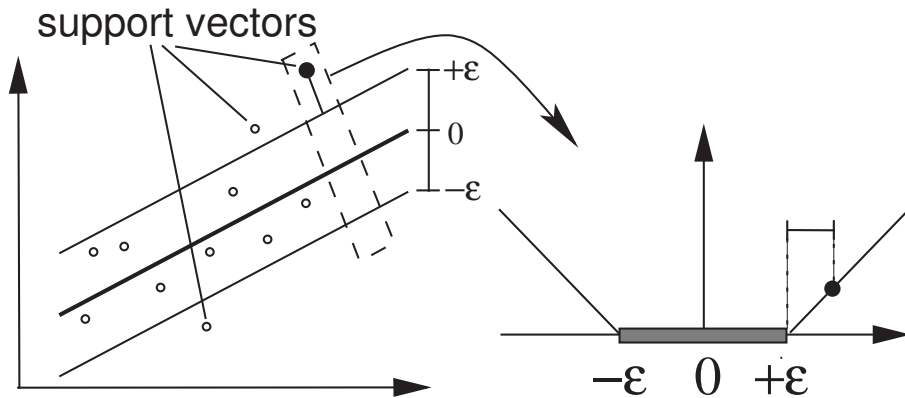


Figure 5.: Utilization of slack variables in ϵ Support Vector Regression. Adapted from Smola and Schölkopf (2004)

cost. In this situation, the problem is formulated as (Smola and Schölkopf, 2004):

$$\begin{aligned}
 & \text{minimize} && \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_i (\zeta_i + \zeta_i^*) \\
 & \text{subject to} && y_i - \mathbf{x}_i \cdot \mathbf{w} - b \leq \epsilon + \zeta_i \\
 & && \mathbf{x}_i \cdot \mathbf{w} + b - y_i \leq \epsilon + \zeta_i^* \\
 & && \zeta_i, \zeta_i^* \geq 0
 \end{aligned}$$

The algorithm tries to minimize both \mathbf{w} and training errors. The selection of C and ϵ determines the trade-off between the maximization of the flatness (i.e., minimization of \mathbf{w}) and the minimization of errors. For instance, a very large ϵ value will create a very flat and meaningless predictor while a very small ϵ value will cause the existence of many nonzero errors. The value of the new instances in regression problems is:

$$b + \sum_{i \text{ is a support vector}} \alpha_i \mathbf{x}_i \cdot \mathbf{a} \tag{18}$$

Unlike classification, the α_i values may be negative for regression. Nonlinear problems are addressed by applying the kernel trick.

In summary, Support Vector algorithms are considered very accurate because they can produce subtle and complex nonlinear decision boundaries (Witten et al., 2011;

Han et al., 2011). Usually, their generalization performance is equal or superior than most methods (Burges, 1998) and they always find a global solution unlike other methods such as NN (Burges, 1998; Han et al., 2011).

However, Support Vector algorithms are slow (Burges, 1998; Han et al., 2011), specially in the training phase. Therefore, their application may be impracticable for very large data sets and limited computational resources. Moreover, there are various hyperparameters that need to be tuned.

2.2.4 *Random Forest*

RF is an efficient ensemble method proposed by Breiman (Breiman, 2001) for classification and regression cases. It is composed by a collection of tree predictors that are created based on independent identically distributed random vectors. The output of a RF model is the most voted class for classification problems or the average prediction value for regression situations.

Decision trees break down problems into increasingly discrete subsets (Turban et al., 2008). They have a tree-like structure constituted by a root followed by internal nodes. Each node is associated with a question and the arcs linked to each node represent all possible answers. Terminal nodes indicate the output selected by the algorithm. Therefore, the output is obtained by answering a sequence of nested questions (Seni and Elder, 2010).

Ensemble methods are usually more accurate than their individual components (Dietterich, 2000). Hansen and Salamon (1990) defend that ensemble is more accurate than any of its components if its individual members are accurate and diverse. Since uncorrelated classifiers frequently assign incorrect labels for distinct instances, it is less likely that an instance is misclassified by the majority of the classifiers. Figure 6 presents a linearly separable problem using an individual decision tree and an ensemble of decision trees.

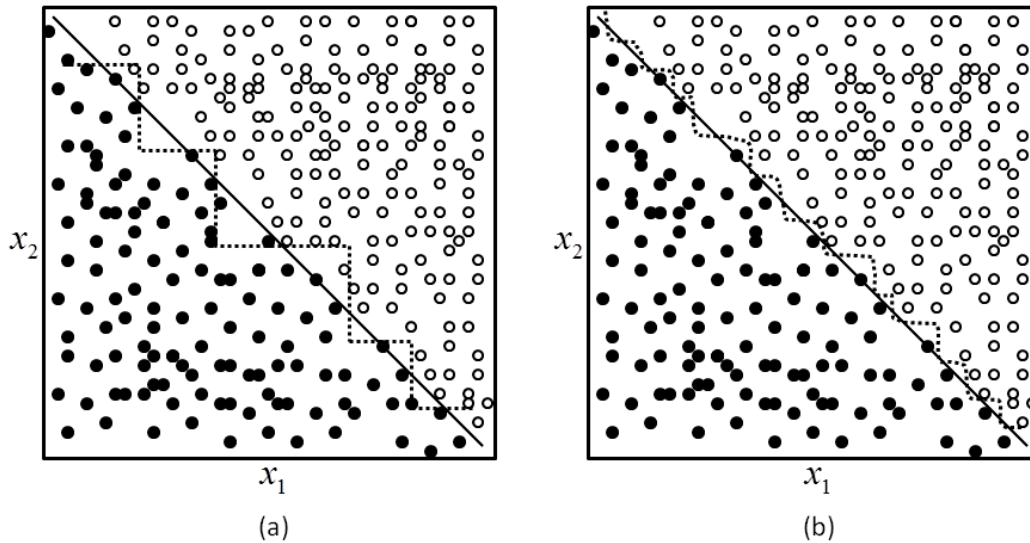


Figure 6.: Decision boundary using an individual decision tree (figure in the left) and an ensemble of decision trees (figure in the right). Adapted from Seni and Elder (2010)

The individual decision tree represented by (a) is ineffective to approximate the true decision boundary represented by the straight line. The ensemble represented by (b) produces a much closer decision boundary.

Breiman was influenced by previous studies on random subspace method (Ho, 1998), random split selection (Dietterich, 2000) and geometric feature selection (Amit and Geman, 1997). At each node of the individual decision trees composing RF, the best split is determined within a small subset of features. The Forest-RI approach randomly selects a small set of the original input variables while the Forest-RC produces new features that are linear combinations of some inputs (Breiman, 2001). In the latter approach, each attribute is created by adding a randomly selected subset of inputs multiplied by coefficients that are uniform random numbers on $[-1,1]$. For instance, F linear combinations can be produced from L variables for each node. Forest-RC is particularly useful when there are few attributes. The selection of a substantial part of the existing attributes may produce highly correlated trees. Therefore, the creation of a larger number of features may reduce the correlation between individual trees. For both approaches, the groups of features are chosen independently with the same

distribution for all trees. The best split is found within the subset of input variables (Forest-RI) or within the set of created features (Forest-RC).

The utilization of bagging seems to improve accuracy when random features are applied (Breiman, 2001). In bagging, each training set is sampled with replacement from the original data set. Therefore, some tuples of the original data set may not occur in any training set, whereas others may appear in more than one training set (Han et al., 2011). Thus, each decision tree of RF can be built using a distinct training set.

RF individual trees are grown to maximum size without pruning using Classification And Regression Trees (CART) methodology (Biau, 2012). After a large number of trees is created, the final output of RF is the most popular class (classification) or the average value of each prediction (regression). The strong law of large numbers shows that the generalization error for RF always converges to a limit as more trees are included (Breiman, 2001). Therefore, overfitting is not a severe problem for RF.

Based on Amit and Geman (1997) analysis, Breiman (2001) also finds that the accuracy of RF depends on the quality of the individual trees and the dependence between them. A RF composed by accurate individual trees having small correlation among them should produce good results.

Bagging permits the calculation of ongoing estimates of the generalization error, classifier strength and dependence. Out-of-bag (OOB) estimates are based on the decision trees that do not applied the current instance in their training process. For instance, the OOB estimate of the training instance t_1 is calculated by combining trees that used bootstrap training sets without t_1 . OOB values provide fair estimation compared to the test set error (Genuer et al., 2008). These internal estimates are useful to understand model accuracy and how to enhance it. Moreover, it removes the necessity for a set aside test set (Breiman, 2001).

RF have diverse qualities. They can process a very large number of variables without overfitting and can generate very accurate predictions (Biau, 2012). RF accuracy is comparable to Adaboost (Breiman, 2001; Han et al., 2011). They are relatively robust

to noise and outliers (Breiman, 2001). RF are efficient on large data sets because they take substantially fewer features for each split (Han et al., 2011). They are faster than boosting or bagging (Breiman, 2001). Since each individual tree can be processed independently, RF can be easily parallelized (Han et al., 2011; Breiman, 2001). Moreover, RF are easy to implement and provide helpful internal estimates of strength, dependence, error and variable importance (Breiman, 2001).

However, RF are not easy to interpret. Furthermore, the computational effort to train the model for several small or medium sized data sets is usually larger than a single NN or SVM training because RF applies often a large number of trees.

2.2.5 *Ensembles*

Ensemble methods combine diverse models in order to produce potentially better predictions than their individual components. Indeed, this methodology resembles human behavior in several situations. For instance, it is common to ask for different medical opinions on serious health problems or to read several user reviews when buying an expensive car. The analysis, weighting and combination of these opinions permits a more informed and confident decision process (Polikar, 2006).

The Condorcet Jury Theorem states that the majority vote of a jury is more accurate than its individual votes for binary outcomes if the accuracy of each member is higher than 50% and their votes are independent. Moreover, the accuracy of the majority approaches 100% as the number of its elements approximates to infinity. Hence, an efficient decision process is assured by the combination of the votes of a large jury constituted by members that are slightly more accurate than random. This theorem was formulated to support the theoretical basis for democracy but it is useful to understand ensemble methodology (Rokach, 2010). Thus, according to the Condorcet Jury Theorem, the combination of independent models more accurate than random may produce a stronger ensemble for binary outputs. In fact, the utilization of diversified

models producing uncorrelated predictions is considered an essential condition for the creation of accurate ensembles (Hansen and Salamon, 1990; Polikar, 2006; Rokach, 2010). The rationale is that the combination of predictions (e.g., most voted class) is more efficient if the errors made by individual models are non-correlated (i.e., less concentrated errors).

There are some possible arguments for the application of ensembles. For instance, some models may obtain good performance on training data but have unsatisfactory generalization performance. The combination of the outputs of diverse models by averaging may reduce the risk of selecting an inadequate single model (Dietterich, 2000; Polikar, 2006).

Moreover, ensembles can allow a more efficient utilization of both large and small volumes of data by training different models on partitions of big data sets or on samples of undersized data sets generated by resampling techniques (Polikar, 2006).

Additionally, the usage of various data sets having distinct characteristics (e.g., different sources, heterogeneous features) may recommend the utilization of diverse models. For example, a doctor may order several tests to diagnose a disease. Each test may require the application of a different model. The final decision can be obtained by the combination of the diverse outputs produced by the individual models (Polikar, 2006).

Furthermore, the decision boundary of some problems can be too complex to be adequately implemented by an individual model. An appropriate combination of diverse decisions boundaries generated by individual models may produce a function closer to the true boundary (Dietterich, 2000; Polikar, 2006).

The ensemble systems are composed by two main components. The first component is the construction of a diversified set of models. This diversity can be obtained in various ways such as (Polikar, 2006):

- usage of different training data sets for individual models (e.g., bootstrapping, bagging);
- utilization of different training parameters for individual models;

Chapter 2. Background

- application of different types of ML models (e.g., NN, SVM, MR, decision trees).

The second component is the combination of the outputs generated by the individual models. There are trainable combination rules that are created by a separate training algorithm and non-trainable rules that do not require separate training. Moreover, there are different rules to combine class labels and continuous outputs. The combination of class labels can be obtained by diverse approaches such as different forms of majority voting. For instance, the ensemble may select the label predicted by all classifiers (unanimous voting), by more than half classifiers (simple majority), by the highest number of classifiers (plurality voting) or adopt a weighted majority voting by assigning different weights to the decisions of different classifiers (e.g., higher weights to more accurate classifiers). The combination of continuous outputs can be achieved in several ways such as the average (simple or weighted), minimum, maximum or median of all outputs.

The best combination and ensemble method depends on the problem. There are no best approach for all situations. Some selection criteria are (Rokach, 2010):

- the accuracy obtained by the ensemble;
- the computational cost for creating the ensemble and for classifying new instances;
- the interpretability of the ensemble;
- the scalability of the ensemble;
- the software availability (e.g., high availability means that there are diverse software alternatives for the ensemble method).

For instance, the average of all outputs (EA) is widely adopted in regression tasks due to its simplicity and satisfactory performance over a broad range of problems (Polikar, 2006). In this work, we adopt this simple ensemble method, by averaging the predictions of four individual ML models (MR, NN, SVM and RF).

2.3 SENTIMENT ANALYSIS

2.3.1 Background

SA is the computational analysis of opinion, sentiment and subjectivity in text (Pang and Lee, 2008). These systems aim to automatically extract and classify sentiment, opinions, emotions, attitudes and appraisals toward specific targets (e.g., events, entities, individuals, topics, issues, attributes) from textual contents (Liu and Zhang, 2012). This field of study is also frequently known as "Opinion Mining" (Pang and Lee, 2008; Liu and Zhang, 2012).

Knowing people's opinion has always been an important piece of information for decision-making processes (Pang and Lee, 2008). For example, companies always want to know consumer opinions about their products to be able to improve them and to perform adequate marketing actions. Potential buyers also want to analyze opinions of actual customers before they purchase a product. Political voters usually seek and share political information about elections and may publish their political views.

Social Media platforms (e.g., reviews, forums, blogs and social networks) have enabled an explosion of contents containing opinions regarding several topics. For instance, there are 1.18 billion daily active users in Facebook¹, 1 billion visits monthly to webpages with embedded Tweets² and 236 million monthly active users in chinese service Weibo³. These users spend a significant part of their time on social media services (Fan and Gordon, 2014). Therefore, they reveal information about various aspects of their lives in these services.

The huge amount of opinionated text in these platforms is a valuable source of opinions of a representative community of users that is very useful for organizations and individuals. Nevertheless, the extraction of these opinions is extremely difficult

¹ <http://newsroom.fb.com/company-info/> [Accessed on 9 January 2017]

² <https://about.twitter.com/company> [Accessed on 9 January 2017]

³ <http://ir.weibo.com/phoenix.zhtml?c=253076&p=iro1-newsArticle&ID=2145407> [Accessed on 9 January 2017]

Chapter 2. Background

for humans. The identification and summarization of important information in large quantities of data is very challenging for the average reader and human analysis can be biased (Liu and Zhang, 2012). These limitations can be overcome by SA systems that mine large amounts of opinionated content and automatically extract and summarize the opinions about a specific topic.

2.3.2 *Opinion Representation*

An opinion can be characterized by diverse features that may be applied in the subsequent analysis. There are five features that are often extracted (Liu and Zhang, 2012):

- **Entity:** Target object that has been evaluated (e.g. product, service, person, event, organization, topic).
- **Aspect:** The entity attribute that has been measured (e.g., component, service, price).
- **Orientation:** The opinion orientation about the aspect of the entity. It can have diverse categories (e.g. positive, negative or neutral) or be expressed with continuous sentiment values.
- **Opinion Holder:** The entity that has expressed the opinion.
- **Time:** The time when the opinion was expressed.

For example, the opinion of the sentence "The new iPad's processor is fast." can be expressed by the quintuple (Entity: iPad, Aspect: processor, Orientation: positive, Opinion Holder: review author, Time: time of the review). The extraction of these features enables the transformation of unstructured text into structured data and permit a more complete and effective knowledge discovery phase.

SA systems may produce discrete or continuous sentiment values. The majority of them use sentiment polarity (Tsytsarau and Palpanas, 2012) (i.e., positive and negative

categories) but some studies apply other scales (e.g., positive, neutral, negative categories; one to five stars), the six basic emotions (Ekman et al., 2013) (i.e., anger, disgust, fear, happiness, sadness, and surprise) or a continuous sentiment representation. The SA model can assign neutral sentiment values to non-opinionative phrases when using these sentiment representations (Tsytsarau and Palpanas, 2012).

We can also have opinions comparing two or more entities. It can indicate a preference of the opinion holder and express differences or resemblances between these entities. The set of five features that describe regular opinions is insufficient to characterize comparative opinions. These opinions may be defined by six features: entity 1, entity 2, shared aspects, preferred entity, opinion holder and time (Liu and Zhang, 2012).

2.3.3 *Classification Levels*

SA can be performed at different levels. The document-level sentiment classification assigns a sentiment value to the whole document. It assumes that the document expresses opinions on a single entity from a single opinion holder (Liu and Zhang, 2012). Thus, it is not appropriated to evaluate and compare diverse entities neither to identify the sentiment regarding multiple entities and aspects mentioned in the document. Customer reviews are well suited to this type of classification because they are generally written by a single author and are about a single item. Moreover, these reviews are a good data source for supervised learning because they are often classified by their own authors. However, document-level sentiment classification is too coarse for some applications.

The sentence level classification may provide a more adequate analysis in some situations because it gives the sentiment of a larger number of information units. However, it is still not adequate for complex sentences that include opinions on multiple aspects.

Chapter 2. Background

Sentence level is appropriate for simple sentences that contain a single opinion from a single holder (Liu and Zhang, 2012).

In many cases, the document level and sentence level classification do not supply the necessary detail. For example, a positive document or sentence does not mean that the author has positive opinions on all mentioned entities or aspects. For instance, the sentence "I like this cell phone but I think the screen is too large and the battery life is small" has an overall positive sentiment but indicates a negative opinion about the screen dimension and battery life. This detailed information could be extremely valuable for decision-making and to product improvement. Aspect-based SA seeks to assign the sentiment orientation of all aspects and entities. Unlike the other levels, the aspect level classification is able to discover all five items that characterize each opinion: entity, aspect, opinion orientation, opinion holder and time (Liu and Zhang, 2012). It is a more complex task that requires deeper NLP capabilities to create a richer set of results.

Furthermore, some textual contents compare distinct entities. These comparative opinions are different from regular opinions, because they involve more entities and have other semantic meanings and syntactic forms. Additional features may be necessary such as the identification of a second entity, the preferred entity and the shared aspects.

2.3.4 Methodology

Research studies about SA have been applying several distinct approaches. For instance, diverse ML algorithms (e.g., SVM, NN, Maximum Entropy (ME)) have been experimented, different data sources (e.g., reviews, blogs, microblogs) have been explored and various domains (e.g., movies, products or politics) have been using SA. These approaches can be grouped in three different types: ML based, lexicon based and a hybrid approach combining both ML and lexicon based methods.

The ML approach applies algorithms capable to learn the SA model from a large training corpus and then use it to assign sentiment values (e.g., categories, continuous values) to each new instance. Most of these approaches are supervised, making use of available labeled data. For instance, a considerable part of research in this topic uses large corpus of product reviews that are rated by their own authors. In ML models, each document is usually represented by a vector of features such as term frequencies or syntactic features. Diverse ML methods (e.g., NB, ME, SVM, NN, Decision Trees) are used in SA problems.

Unsupervised SA algorithms are mostly dependent on opinion lexicons or dictionaries. The lexicon based approach uses a list of known opinion sentiment terms to determine the sentiment value. Most lexicon methods aggregate the sentiment values of lexicon entries occurring in each document. A common formula to compute the sentiment value is (Tsytsarau and Palpanas, 2012):

$$S(D) = \frac{\sum_{w \in D} S_w \cdot weight(w) \cdot modifier(w)}{\sum weight(w)} \quad (19)$$

where S_w is the lexicon sentiment value for the document word w , $weight(w)$ is a measure that may affect sentiment based on some criteria (e.g., distance from specific words) and $modifier(w)$ addresses situations that may cause alterations in sentiment such as negation or intensity words. The assigned sentiment category is based on this calculated value.

There are various available lexicons that can be easily and promptly applied in SA (e.g., MPQA (Wilson et al., 2005), SWN (Baccianella et al., 2010), GI (Stone et al., 1966)). However, other SA works also produce their own lexicons (e.g., Mohammad et al., 2013; Kiritchenko et al., 2014). Some publicly available lexicons were manually created by experts but this approach is too much expensive and laborious to be applied regularly. The automatic creation is faster and generate more entries but it is usually less accurate. The automatic approach explores text corpora or existing dictionaries (e.g., WordNet) and thesaurus. The extraction of opinion words and the calculation of their sentiment

value is mainly based on syntactic relations, POS information, statistical measures (e.g., PMI) or semantic relations (e.g., synonyms, antonyms). Other studies iteratively identify and classify opinion terms by applying semantic and syntactic relations on a short seed list of classified words.

Lexicon based methods permit an easy unsupervised rule-based classification that is computationally less complex than ML methods and does not depend on labeled data. Moreover, lexicon based methods can easily produce continuous sentiment values. On the other hand, models based on lexicons frequently have low recall values because there are often documents without any lexicon term (Martínez-Cámara et al., 2014). ML models usually have more domain adaptability (Tsytsarau and Palpanas, 2012). There are some studies comparing the performance of ML and lexicon based algorithms. The results are not consensual. For instance, Chaovalit and Zhou (2005) and Annett and Kondrak (2008) obtained similar precision values for both approaches while Gindl and Liegl (2008) found a superiority of the ML methods over the lexicon based methods. The utilization of lexicons to produce features (e.g., identify opinion terms, calculate a sentiment score) for ML models is also very common.

An alternative approach is a hybrid combination of ML and lexicon based algorithms to explore the advantages of both methods (Medhat et al., 2014; Martínez-Cámara et al., 2014). For instance, Prabowo and Thelwall (2009) apply a procedure composed by a sequence of lexicon-based and ML classifiers. Each classifier processes the documents that were not classified by the previous method. The utilization of lexicon-based algorithms before a SVM classifier improved the performance compared to a stand-alone SVM classifier. This approach complemented the high precision of lexicon-based methods with the higher recall of ML methods.

A significant part of SA models uses BOW representations. In this approach, each document is simply represented by its words, thus ignoring grammar and word order. Some other studies defend a different representation using concepts to characterize textual contents. The concept-based approach uses ontologies or semantic networks to perform semantic text analysis. This method intends to extract conceptual information

by applying large semantic knowledge bases. It may enhance other SA approaches by performing an adequate analysis of sentiment on non affective expressions that are associated to concepts conveying emotion (Cambria et al., 2013). Concept-level SA systems have been applied in various areas such as e-health (Cambria et al., 2012), movies (Mudinas et al., 2012; Poria et al., 2014) or smartphones (Kontopoulos et al., 2013).

Other approaches attempt to address the principle of compositionality. This principle refers that the semantic value of a complex expression is dependent on the meaning of its components and the syntactic rules combining them (Montague, 1974). The BOW methods may be ineffective in some situations because they do not represent sentence components interactions. For example, the sentence "I do not hate it" has two usual negative terms (i.e., "not", "hate") but the overall sentiment is not negative. Sentiment composition algorithms try to handle compositional effects such as polarity reversal or sentiment propagation by using rules (Moilanen and Pulman, 2007; Choi and Cardie, 2008; Heerschop et al., 2011; Neviarouskaya et al., 2010) or ML methods (Socher et al., 2013).

Table 1 presents information about a comprehensive set of important SA papers. For each study, it is presented the components of the applied algorithm (e.g., SVM, NN, Lexicon), the different data sources (e.g., Twitter, Reviews, Blogs), the domain of the study (e.g., Products, Finance, Movies) and the main SA approach type (i.e., ML, Lexicon based or Hybrid).

ML algorithms are heavily used in SA problems. For instance, 68% of the analyzed papers apply ML methods in their SA algorithms. The usage of lexicons is also very common in SA, as it was already identified in previous SemEval competitions (Rosenthal et al., 2014, 2015a; Nakov et al., 2016). Indeed, nearly half of the studies use lexicon based or hybrid approaches. Moreover, some ML approaches apply opinion lexicons to select features for their models (e.g., Hagenau2013, Li2013). The combination of ML and lexicons methods (i.e., hybrid approaches) already have a reasonable utilization (17%), probably to exploit some potential complementary characteristics of both

Chapter 2. Background

Table 1.: Summary of SA literature

Author Algorithms*	Data Source	Domain	Type
(Spertus, 1997) DT	Comments	Politics	ML
(Morinaga et al., 2002) Lexicon, Statistics	Web pages	Products	Lexicon
(Dave et al., 2003) Statistics, SVM, NB	Reviews	Products	ML
(Laver et al., 2003) Statistics	Party Manifestos	Politics	ML
(Liu et al., 2003) Lexicon	Knowledge bases	General	Lexicon
(Yi et al., 2003) Lexicon	Reviews	Products	Lexicon
(Yu and Hatzivassiloglou, 2003) NB, Statistics, Lexicon	News	General	Hybrid
(Hu and Liu, 2004) Lexicon	Reviews	Products	Lexicon
(Kim and Hovy, 2004) Lexicon, Statistics	Text Corpus	General	Lexicon
(Kudo and Matsumoto, 2004) DT	Reviews, News	Products	ML
(Alm et al., 2005) P, Lexicon	Stories	Literature	ML
(Chaovalit and Zhou, 2005) Statistics, Lexicon, P	Reviews	Movies	ML and Lexicon
(Pang and Lee, 2005) SVM	Reviews	Movies	ML
(Goldberg and Zhu, 2006) Graphs	Reviews	Movies	ML
(Ku et al., 2006) Lexicon, Statistics	News, Blogs	General	Lexicon
(Leung et al., 2006) Lexicon	Reviews	Movies	Lexicon
(Taboada et al., 2006) Lexicon	Reviews	Products	Lexicon
(Thomas et al., 2006) SVM	Debate transcripts	Politics	ML
(Devitt and Ahmad, 2007) Graphs, Lexicon, Statistics	News	Finance	Hybrid
(Godbole et al., 2007) Lexicon	Blogs, News	General	Lexicon
(Mei et al., 2007) Statistics	Blogs	General	ML
(Moilanen and Pulman, 2007) Lexicon	News	General	Lexicon
(Annett and Kondrak, 2008) SVM	Blogs	Movies	ML
(Choi and Cardie, 2008) Lexicon, SVM	News	General	Hybrid
(Choi et al., 2009) Lexicon, Statistics, C	News	Newspapers	Hybrid
(Go et al., 2009) NB, ME, SVM	Twitter	General	ML
(Lerman et al., 2009) SVM	Reviews	Products	ML
(Lin and He, 2009) LDA, Lexicon	Reviews	Movies	Hybrid
(Melville et al., 2009) Lexicon, NB	Blogs, Reviews	Products, Politics, Movies	Hybrid
(Miao et al., 2009) Lexicon, Statistics	Reviews	Products	Lexicon
(O'Hare et al., 2009) NB, SVM	Blogs	Finance	ML
(Prabowo and Thelwall, 2009) Lexicon, Statistics, SVM	Reviews, Comments	Movies, Products	Hybrid
(Asur and Huberman, 2010) P	Twitter	Movies	ML
(Bifet and Frank, 2010) Statistics, NB, DT	Twitter	General	ML
(Neviarouskaya et al., 2010) Lexicon	Message Board	General	Lexicon
(O'Connor et al., 2010) Lexicon	Twitter	Politics, Economy	Lexicon
(Pak and Paroubek, 2010) NB	Twitter	General	ML
(Tumasjan et al., 2010) Lexicon	Twitter	Politics	Lexicon
(Bai, 2011) Graphs, MKB	Reviews, News	Movies, General	ML
(Bermingham and Smeaton, 2011) NB	Twitter	Politics	ML
(Bollen et al., 2011) Lexicon	Twitter	Finance	Lexicon
(Heerschop et al., 2011) Lexicon, GA	Reviews	Movies	Lexicon
(Jiang et al., 2011) SVM, Graphs	Twitter	General	ML
(Marcus et al., 2011) NB	Twitter	Events	ML
(Maynard and Funk, 2011) Lexicon	Twitter	Politics	Lexicon
(Thelwall et al., 2011) Lexicon	Twitter	Events	Lexicon
(Xu et al., 2011) CRF	Reviews	Products	ML
(He and Zhou, 2011) Lexicon, Statistics	Reviews	Movies, Products	Lexicon
(Zhang et al., 2011a) Lexicon, Statistics, SVM	Twitter	General	Hybrid
(Zirn et al., 2011) Lexicon, MLN	Reviews	Products	Hybrid
(Cambria et al., 2012) Statistics, Graphs, C	Questionnaires	Healthcare	ML
(Duric and Song, 2012) ME, Statistics	Reviews	Movies	ML
(Hu et al., 2012) Lexicon	Reviews	Products	Lexicon
(Kang et al., 2012) NB, Lexicon	Reviews	Restaurants	Hybrid
(Mudinas et al., 2012) Lexicon, SVM	Reviews	Software, Movies	Hybrid
(Hagenau et al., 2013) SVM	News	Finance	ML
(Kontopoulos et al., 2013) FCA, OL, Lexicon	Twitter	General	Hybrid
(Lamos et al., 2013) BLR	Twitter	Politics	ML
(Li and Li, 2013) SVM	Twitter	Brands, Products	ML
(Mohammad et al., 2013) SVM, Lexicon, Statistics	Twitter, SMS	General	ML
(Moraes et al., 2013) SVM, NN, NB	Reviews	Movies, Products	ML
(Rui et al., 2013) Lexicon, NB, SVM	Twitter	Movies	ML
(Socher et al., 2013) NN	Reviews	Movies	ML
(Hutto and Gilbert, 2014) Lexicon	Twitter, Reviews, News	General	Lexicon
(Kiritchenko et al., 2014) SVM, Lexicon, Statistics	Twitter, SMS	General	ML
(Poria et al., 2014) SVD, SVM, NN	Reviews	Movies, Products	ML
(Tang et al., 2014b) NN, SVM	Twitter	General	ML
(Tang et al., 2014a) NN, SVM, Lexicon	Twitter	General	ML

* SA algorithm components: BLR – bilinear regression, C – clustering, CRF – conditional random fields, DT - decision trees, FCA - formal concept analysis, GA – genetic algorithm, LDA – latent Dirichlet allocation, MKB – Markov blanket, MLN – Markov logic networks, ME – maximum entropy, NB – naive bayes, NN – neural networks, OL – ontology learning, P – sentiment analysis product, SVD – singular vector decomposition, SVM – support vector machines

methods (e.g., higher recall for ML, easy unsupervised classification of lexicon based models). For example, lexicon based methods may produce an initial labeled training data set for ML classifiers. Figure 7 presents the number of studies for each approach type.

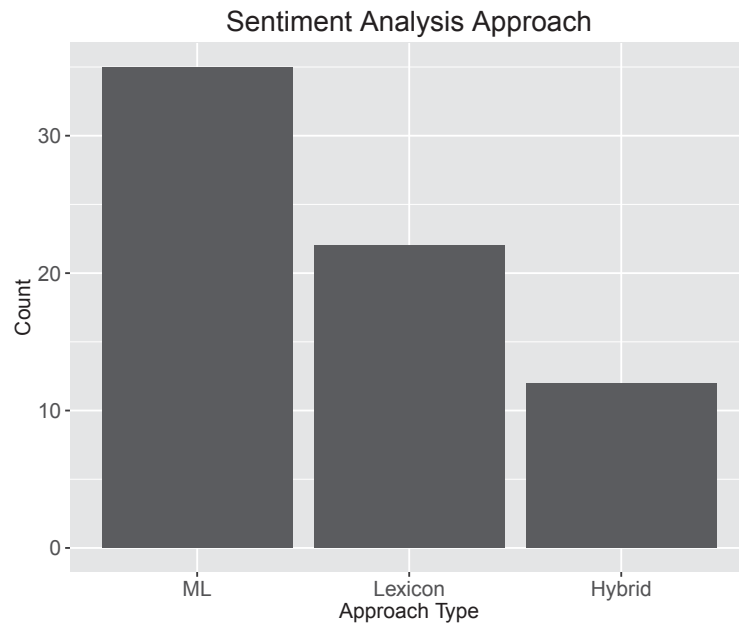


Figure 7.: SA approaches

SA algorithms are composed by several different components. In these studies, we identified 19 distinct algorithm components. Nevertheless, four elements are much more applied than average: lexicons, SVM, statistical measures and NB. Most components are scarcely utilized (e.g., 47% of the components are only used in one study). SVM and NB are also among the most popular classifiers in SA tasks of SemEval competitions (Rosenthal et al., 2014, 2015a) alongside ME, Conditional Random Fields and linear regression. An increasing utilization of deep learning is also evident in the most recent research studies and in the last editions of SemEval. There was a dominance of deep learning methods (e.g., convolutional and recurrent NN) in SemEval2016 (Nakov et al., 2016) and deep NN systems won several SA subtasks of SemEval2015 (Rosenthal et al., 2015a). The usage of the algorithm components in the analyzed papers is displayed in Figure 8.

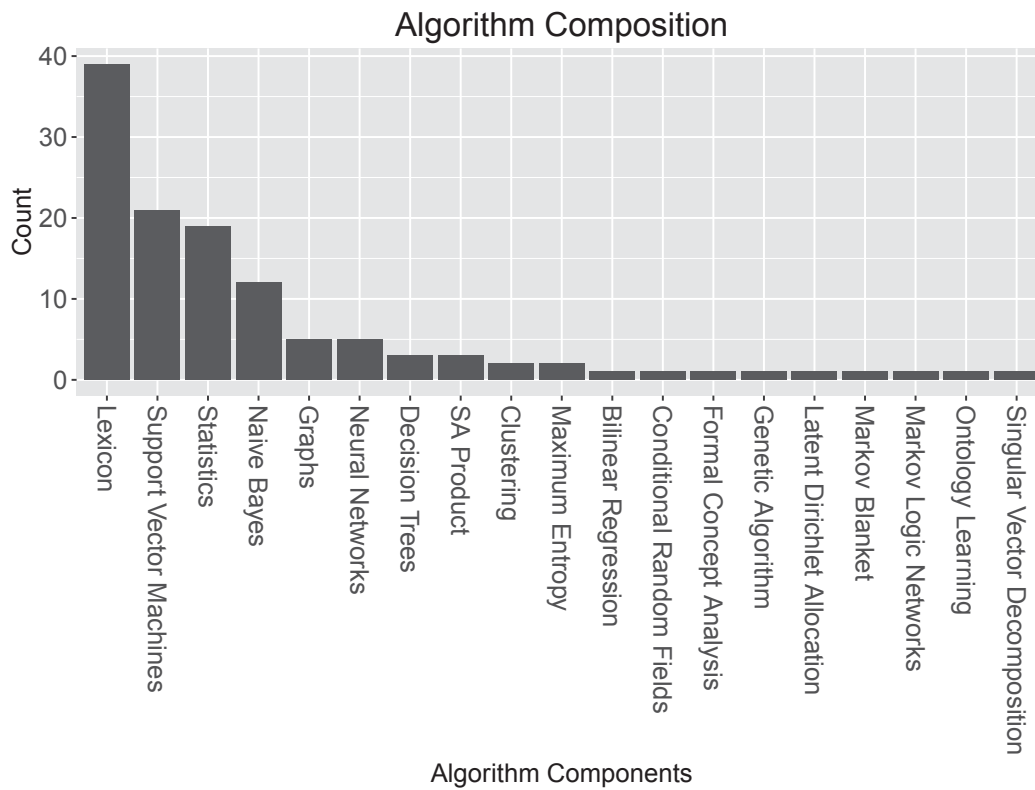


Figure 8.: SA algorithm components

2.3.5 Features

The features used in SA models are quite varied. The term frequencies are applied in a significant part of SA models. They are composed by individual words or n-grams (i.e., contiguous sequence of words) and the respective number of occurrences. Since some uninformative terms may also have high term frequency (e.g., "the", "a"), several SA models use the Term Frequency - Inverse Document Frequency (TFIDF) weighting scheme. These values decrease proportionally to the number of documents containing each term. Thus, terms appearing frequently in most documents have their values adjusted (e.g., the word "the" has an inferior relative TFIDF value than term frequency). The term presence is also widely applied in SA. This binary representation indicates whether a term is present or not. The term presence value is zero for absent

terms and it is one for terms occurring in the document, regardless of the number of occurrences.

The presence or frequency of a set of opinion terms is usually applied in SA. The selection of the best set of keywords is non-trivial (Pang and Lee, 2008). These words or n-grams have higher expected value for SA and they can be used in unsupervised lexicon-based methods or as features for ML methods. There are diverse available opinion lexicon that can be easily applied in SA models such as MPQA (Wilson et al., 2005), SWN (Baccianella et al., 2010) or GI (Stone et al., 1966). These lexicons can have terms associated to various affect categories, sentiment orientations or sentiment continuous scores. For instance, lexicons are usually heavily applied in the "Sentiment Analysis in Twitter Task" of SemEval competitions (Rosenthal et al., 2014, 2015a; Nakov et al., 2016).

The position of a term in a document (e.g., begin, middle or end of a document) may also influence the sentiment classification. Therefore, some SA algorithms include the position information in their models (Pang et al., 2002). The word shape, the presence of elongated words (e.g., "sooo"), punctuation (e.g., presence of exclamation and question marks) and specific terminology (e.g., presence of emoticons or hashtags in tweets) are also features frequently applied in SA models (Rosenthal et al., 2014, 2015a).

The POS are categories corresponding to certain syntactic functions (e.g., noun, adjective, verb, adverb, pronoun). They may be useful for word sense disambiguation or to extract terms that may have higher sentiment value (e.g., adjectives (Hatzivassiloglou and McKeown, 1997; Wiebe, 2000; Hu and Liu, 2004)). Other syntactic features such as word dependency relations are also used in various SA models (e.g., Kudo and Matsumoto, 2004), particularly in short text documents.

Opinion orientation is usually affected by negation words or expressions. This influence may even be decisive in shorter documents. However, the impact of negation may vary. For instance, sentiment may be reversed (e.g., don't like) or just decreased (e.g., not frightening) by negation. Some SA algorithms try to address negation by

Chapter 2. Background

using different features (e.g., attaching NOT to words near negation terms) or rules (e.g., reverse sentiment polarity). Negation handling is considered important for the performance of SA models (Rosenthal et al., 2015a).

The textual and semantic relations between sentence segments can also influence the overall sentiment value. The diverse discourse relationships (e.g., causation, contrast, explanation) have distinct impact on sentiment. For instance, the existence of terms such as "but", "however", "in contrast" may imply a twist in the global sentiment of the document. The BOW representation is inadequate to model this type of information. An approach to include discourse structure is to model the overall sentiment as a trajectory of local sentiments (e.g., Mao and Lebanon, 2006).

A feature that has been increasingly applied in SA is word embeddings. They correspond to representations of words or expressions by vectors of real numbers. The word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) are two popular unsupervised algorithms that produce word embeddings. For instance, it was widely used in the track "Sentiment Analysis in Twitter Task" of the last edition of the competition SemEval (Nakov et al., 2016).

The reduction of the dimensionality of the data allow the acceleration of the SA process, the economization of computer resources and may even improve SA results. To select features from the original set, there are several methods. The removal of stop words permits the elimination of diverse common uninformative words. Stemming and lemmatization group sets of related words into a common base form by removing some word endings or by providing the lemma. Statistical methods (e.g., Principal Component Analysis, Latent Semantic Indexing) can also be used to produce a smaller set of features.

2.3.6 *Lexicons*

The sentiment lexicon is considered a key element for SA (Feldman, 2013). The utilization of lexicons permits the execution of effective unsupervised approaches (Turney and Littman, 2003; Hatzivassiloglou and Wiebe, 2000) and provides high quality features for supervised SA (Choi et al., 2009; Rosenthal et al., 2015b). Moreover, lexicons can be applied to diverse tasks such as SA, opinion retrieval (Ounis et al., 2008) or opinion question answering and summarization (Dang and Owczarzak, 2008), and they can be applied to diverse domains such as finance (Loughran and McDonald, 2011), electronic products (Hu and Liu, 2004) or the movie industry (Kennedy and Inkpen, 2006).

The creation of opinion lexicons is an important topic that has been studied for some time under two main approaches: manual and automatic creation. Manual creation is the most labor intensive and expensive approach because it requires experts to manually classify the sentiment value of each term. MPQA subjectivity lexicon (Wilson et al., 2005) and GI (Stone et al., 1966) are two important examples of this methodology. Automatic creation requires much less human effort and allows for the faster inclusion of a larger set of lexical items. However, this is often achieved at the expense of accuracy.

There is substantial literature about the automatic construction of lexicons. Many of these studies apply text corpora for this procedure. Hatzivassiloglou and McKeown (1997) extracted conjoined adjectives from a large Wall Street Journal corpus and produced a list of adjectives labeled as positive or negative. Using an initial set of adjectives with predetermined orientation labels, they developed a supervised learning algorithm to assign the sentiment polarity. First, they applied a log-linear regression model to determine if each pair of conjoined adjectives had the same or different orientations. Then, a clustering algorithm was used to divide the adjectives into two different positive and negative sets. Wiebe (2000) extracted subjective adjectives from corpora by applying a method for clustering words based on distributional similarity (Lin, 1998) and another method to compute polarity and gradability (Hatzivassiloglou

and McKeown, 1997). Turney and Littman (2003) tested two co-occurrence measures, PMI and Latent Semantic Analysis, on the AltaVista Advanced Search engine in order to assign a semantic orientation to each word. The most accurate approach calculated the PMI of each term with pre-classified positive and negative words. A term was considered positive if the sum of its PMI scores with positive words was greater than the total PMI score with negative words, and vice-versa. Qiu et al. (2011) explored diverse syntactic relations between opinion words and targets to iteratively extract further opinion words and targets. An initial seed set of opinion words was expanded and opinion targets were collected by continuously identifying terms having those syntactic associations with already extracted terms. Kiritchenko et al. (2014) generated lexicons from tweets containing specific hashtag words and emoticons. These symbols were used as signals of the message sentiment (positive or negative). The sentiment score of each term was calculated using the PMI measures with positive and negative messages. These authors also produced distinct scores for negated and non-negated segments to properly obtain the sentiment in these contexts.

Lexical databases and thesaurus are also extensively applied in the creation of opinion lexicons. Kamps et al. (2004) calculated the synonymy shortest path on the WordNet database (wordnet.princeton.edu) of adjectives to the words “good” and “bad” and determined their sentiment orientation based on these values. Kim and Hovy (2004) created a system that automatically extracts holders and sentiment of each opinion about a given topic. This system includes a module for computing word sentiment. Synonymy and antonymy relations from WordNet are applied in this module in order to expand a small set of seed words and to calculate the strength of sentiment polarity. Esuli and Sebastiani (2005) applied text classification techniques to the glosses of subjective words in order to determine their sentiment polarity. Mohammad et al. (2009) produced a large lexicon using a set of affix patterns and the Macquarie thesaurus. The sentiment of every thesaurus paragraph was classified using a set of positive and negative words collected utilizing affix patterns. Then, each lexical item assumed the most common sentiment label of paragraphs containing the respective term. Baccianella et al.

(2010) produced the SWN lexicon by automatically calculating sentiment values to all WordNet synsets. First, these synsets were classified by a group of classifiers trained with pre-classified synsets. The different classification results were combined to generate a sentiment score to each synset. In a second phase, two iterative random-walk procedures were executed for the positivity and negativity values. These processes applied a graph with directed links from synsets included in glosses of other synsets. The random walk phase began with the values created in the previous step and finished when the processes had converged. Neviarouskaya et al. (2011) created a lexicon by expanding an initial set of lexicon entries through synonymy, antonymy and hyponymy relations, derivation and compounding.

Other works explore both text corpora and lexical databases. Hu and Liu (2004) utilized consumer reviews and Wordnet to select and classify opinion words associated to frequent product attributes. First, they extracted all adjectives included in the sentences of consumer reviews mentioning those product features. The sentiment orientation was assigned according to their semantic association in Wordnet with a seed list of words with known sentiment orientation. Each adjective assumed the same sentiment of synonyms or the inverse polarity of antonyms. The seed list was iteratively expanded with these newly classified words until no further words had antonyms or synonyms in the list. Takamura et al. (2005) proposed the utilization of a spin model, where each word had a sentiment polarity (positive or negative), to produce an opinion lexicon. They created a lexical network based on the occurrence of terms in glosses of other terms, the synonymy, antonymy and hypernymy relations in thesaurus and some conjunctive expressions in corpus. Then, the mean-field method was applied on the network to determine the semantic orientations. Lu et al. (2011) automatically generated a context-dependent sentiment lexicon by combining the utilization of domain independent lexicons, sentiment ratings of reviews, synonym and antonym relations in Wordnet and linguistic rules.

The utilization of generic lexicons or lexicons associated to other domains may be ineffective to SA on stock market text because sentiment is sensitive to the domain

(Turney and Littman, 2003). For example, the verb “underestimate” has often a negative sentiment but an underestimated stock can constitute an opportunity to buy, thus denoting a positive value within the stock market domain. However, the creation of opinion lexicons for the financial domain has been scant. One of the most popular works in this context is by Loughran and McDonald (2011), who manually created six word lists (i.e., positive, negative, litigious, uncertainty, modal strong and modal weak) from words occurring in at least 5% of a large collection of 10-K documents between 1994 and 2008. In 2014, Mao et al. (2014) presented a procedure to automatically produce a Chinese financial lexicon. A large Chinese news corpus was labeled according to the stock returns. Then, a set of seed words was selected based on the Document Frequency value with news associated with very high or very low returns. The lexicon was expanded by considering the statistical association with seed words and the economic significance of candidate terms. The final lexicon was obtained by an iterative optimization process.

In summary, the majority of the studies applies text corpora (e.g., Hatzivassiloglou and McKeown, 1997; Wiebe, 2000; Hu and Liu, 2004; Takamura et al., 2005; Qiu et al., 2011; Kiritchenko et al., 2014; Mao et al., 2014) and/or existing lexical databases and thesaurus (e.g., Kamps et al., 2004; Kim and Hovy, 2004; Hu and Liu, 2004; Esuli and Sebastiani, 2005; Takamura et al., 2005; Kennedy and Inkpen, 2006; Mohammad et al., 2009; Baccianella et al., 2010; Neviarouskaya et al., 2011; Tufiş and Ştefănescu, 2012). The extraction of opinion words or targets are mainly based on syntactic relations (e.g., Hatzivassiloglou and McKeown, 1997; Qiu et al., 2011), POS (e.g., adjectives (Hatzivassiloglou and McKeown, 1997; Wiebe, 2000; Hu and Liu, 2004)), co-occurrence with terms (e.g., Hu and Liu, 2004) and semantic relations in WordNet (e.g., Kim and Hovy, 2004; Neviarouskaya et al., 2011). The calculation of the sentiment polarity or score is mostly performed by statistical measures (e.g., PMI (Turney and Littman, 2003; Kiritchenko et al., 2014)), text classification of glosses (e.g., Esuli and Sebastiani, 2005; Mohammad et al., 2009; Baccianella et al., 2010), semantic associations (e.g., synonymy, antonymy relations in WordNet (Kamps et al., 2004; Kim and Hovy, 2004; Hu and Liu,

2004; Neviarouskaya et al., 2011)), syntactic relations (e.g., Qiu et al., 2011) and clustering methods (e.g., Hatzivassiloglou and McKeown, 1997; Wiebe, 2000). Only one study produced different sentiment scores for affirmative and negated contexts (Kiritchenko et al., 2014), although applied for generic Twitter messages and thus not specifically adjusted to the stock market domain. Some of these studies extract and classify opinions words simultaneously by iteratively expanding a pre-labeled seed list by semantic or syntactic relations (e.g., Hu and Liu, 2004; Qiu et al., 2011). The newly collected words assume the same or the opposite polarity of the associated term.

Some of these approaches are ineffective for the creation of specialized domain lexicons. Methods based on WordNet or thesaurus produce domain independent lexicons, possibly unadjusted for stock market contents (O’Leary, 2011). Therefore, the utilization of methods such as semantic relations in WordNet and text classification of glosses are unsatisfactory for domain dependent lexicons. The usage of a small group of syntactic relations (e.g., conjunctions (Hatzivassiloglou and McKeown, 1997; Takamura et al., 2005)) does not allow for the selection of a large set of words. Moreover, extracting only adjectives (e.g., Hu and Liu, 2004; Kamps et al., 2004; Hatzivassiloglou and McKeown, 1997) ignores terms with a strong sentiment, such as “love” and “hate” verbs. In addition, the expansion of a seed list of classified words (e.g., Qiu et al., 2011) is less effective than the application of classified text corpora. The collection of words co-occurring with specific terms (e.g., Hu and Liu, 2004) and the utilization of clustering methods (e.g., Hatzivassiloglou and McKeown, 1997; Wiebe, 2000) to assign sentiment polarity become less relevant when using classified domain data.

2.3.7 Data Sources

SA has been applied to diverse data sources. However, most studies use data from two different sources: reviews and Twitter. Rated reviews from web pages (e.g., Amazon, IMDB) are a valuable asset for SA in diverse domains. Authors classify and express

their sentiment about diverse products and services and provide large labeled data sets that are valuable for supervised SA. The summarization of these user reviews permit the extraction of important information for organizations and consumers (e.g., brand reputation, product defects). Twitter is an important data source that has been increasingly used for SA. Microblogging data has some characteristics that may add value to SA but also implies some challenges. For instance, their high posting frequency permits the creation of high periodicity indicators. However, microblogging data imply some difficulties such as the informal linguistic style (e.g., abbreviations, jargon, specific terminology), the frequent incorrect use of grammar and the lack of context in many messages (Martínez-Cámara et al., 2014). Therefore, microblogging data require specific pre-processing operations such as the removal of Uniform Resource Locator (URL), replacement of emoticons, word normalization, abbreviation substitution and punctuation elimination (Rosenthal et al., 2014).

News and blogs are less common sources. These contents differ from microblogging data because they usually have a more formal writing and a more correct utilization of grammar. Therefore, the application of standard NLP tools for frequent pre-processing tasks (e.g., stemming, POS tagging) is more effective than for microblogging data. These documents are also much longer than microblogging messages, so they have more context to analyze. However, the posting frequency is much lower. Thus, the aggregated sentiment values may be less representative of the target community (e.g., consumers, investors).

Several other sources are applied in literature (e.g., message board, comments, debate transcripts, party manifestos, SMS, web pages) but their utilization is very scarce. Figure 9 shows the distribution of data sources for SA.

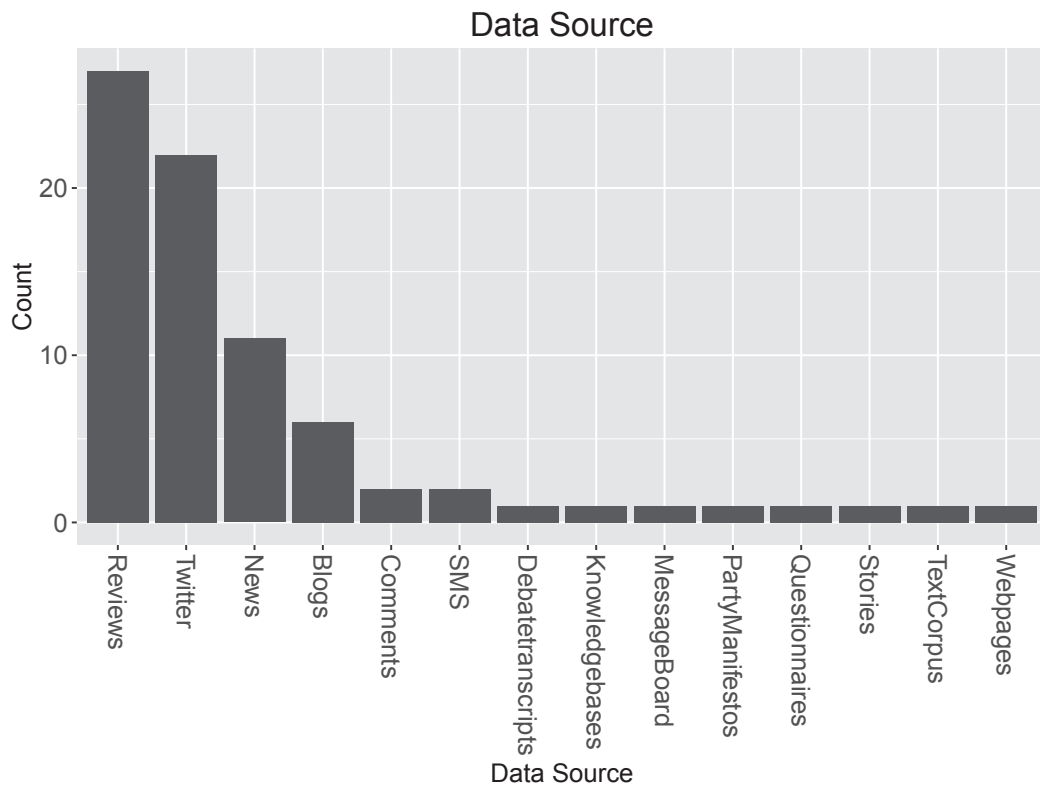


Figure 9.: SA data sources

2.3.8 Applications

The summarization of user reviews is an important domain of SA. The development of automatic tools that extract online reviews and summarize their information can be very useful to understand consumer preferences, identify product qualities and defects, assess brand reputation and recognize user needs. A considerable part of research in SA (25% of the analyzed papers) is related with product reviews (e.g., Hu and Liu, 2004; Dave et al., 2003; Turney, 2002). These systems can be used for diverse purposes such as sales prediction, advertising strategy, reputation management or to support consumer decisions. The analysis of product reviews has been utilized in different sectors such as restaurants (Kang et al., 2012) or software (Mudinas et al., 2012).

Chapter 2. Background

The sentiment classification of movie reviews has been extensively studied (e.g., Pang and Lee, 2005; Asur and Huberman, 2010; Turney, 2002). User reviews in movie webpages (e.g., IMDB) provide an important data set to assess the public opinion about specific movies. For example, this information can be used to predict box-office revenues (Asur and Huberman, 2010).

In politics, SA can be applied for both political parties or voters. While the assessment of voters' opinion is useful for politicians (e.g., O'Connor et al., 2010; Lampos et al., 2013), the extraction of politicians' opinions is important to support voters' decisions (e.g., Laver et al., 2003; Thomas et al., 2006). SA is also used in eRulemaking by analyzing the public opinion about the proposed policies and regulations in order to support the government agencies' decisions (e.g., Kwon et al., 2006).

The assessment of investors or consumer sentiment is also regularly performed for finance (Bollen et al., 2011; Hagenau et al., 2013; O'Hare et al., 2009) and economy (O'Connor et al., 2010). For instance, SA can be used to create alternative consumer sentiment measures (O'Connor et al., 2010) or to produce investor sentiment indicators that can be applied to forecast stock market variables (Bollen et al., 2011).

In healthcare, SA can be applied to evaluate patients health and emotional status by analyzing questionnaires and text descriptions (Cambria et al., 2012). Figure 10 presents the identified domains in SA literature.

SA can also be an important component of other systems. For instance, recommendation systems may benefit from the classification of users opinion by avoiding the recommendation of items having negative feedback (Terveen et al., 1997). Advertising systems may be enhanced by detecting the sentiment towards a product. For example, ads should be avoided when the sentiment is negative or be placed when the sentiment is positive. The identification of hostile messages can be useful for diverse types of communication (e.g., email) (Spertus, 1997). Question answering may improve by properly processing opinion-oriented questions (e.g., Stoyanov et al., 2005) or including the public sentiment about an entity in their answers (e.g., Lita et al., 2005).

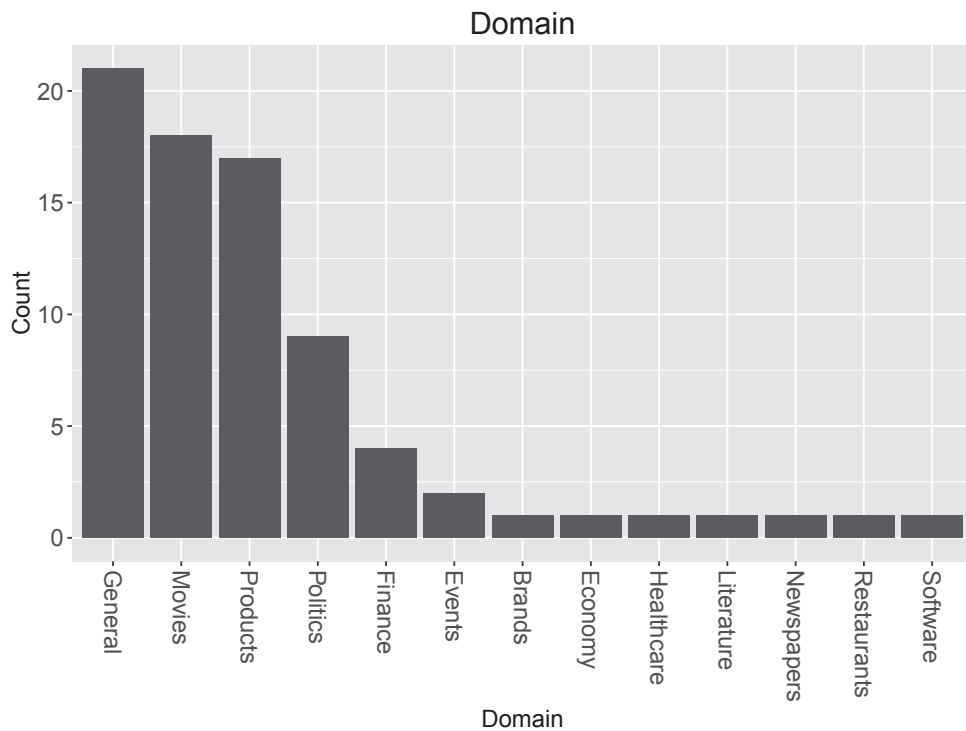


Figure 10.: Domains applying SA

2.4 FINANCIAL MARKETS: BASIC CONCEPTS

2.4.1 Background

Finance can be roughly defined as the field of study associated to the management of financial resources for different entities such as corporations, sovereign states, families or individuals. It involves the allocation of assets over time by analyzing the risk and return of the several available alternatives in order to decide which one maximizes the investor utility. For instance, corporations have to decide which investments should be taken and how they should be funded (e.g., debt or equity). Moreover, families need to analyze diverse banking products (e.g., savings accounts, loans, credit cards), insurance (e.g., life, health), retirement products or investments (e.g., stock market, bonds).

Chapter 2. Background

Financial services such as banks, credit-card companies, insurance companies or stock brokerages provide instruments to satisfy the needs of diverse economic entities. For example, banks can help to match the activities of lenders and borrowers by accepting deposits from lenders and lending those deposits to borrowers. Financial markets aggregate the activities of diverse agents by enabling the transaction of securities and commodities. For instance, corporations can raise capital by selling their securities in the primary market and investors can then sell and buy securities of those corporations in secondary markets.

Nowadays, large volumes of financial data are available to be used by financial entities in order to improve decision making in diverse financial operations. The utilization of ML and DM methods on these data sets may allow the extraction of important information for different objectives such as credit approval, bankruptcy prediction or fraud detection. For instance, the analysis of personal information and previous behavior of a large set of credit applicants may allow a better decision for credit approval. Also, anomaly detection procedures may permit a better identification of fraudulent behavior. This analysis may provide instant results that can be explored nearly in real time.

In the stock market, the application of ML is also very useful to identify investment opportunities. For example, these models can be included in stock market applications that alert investors for these identified opportunities or even be used by algorithms that automatically determine and execute trading strategies. Indeed, algorithmic trading has been applied to identify trading opportunities and place the respective orders at very fast speeds across various markets. These methods permit a rapid analysis of several sources of data and a rigorous, unbiased and non-emotional evaluation of relevant information.

The recent explosive growth of social media utilization also enabled the exploration of large volumes of data describing diverse aspects of their users' lives. Thus, social media data can be used to obtain a better understanding of several financial agents that can be useful for diverse financial operations. For instance, financial applications such as Thomson Reuters Eikon or Bloomberg provide indicators extracted from social

media. Moreover, a deeper perception about the quotidian live of individuals can add important information for tasks such as credit approval or fraud detection.

In summary, the exponential increase of available data and the improvement of computational resources originate valuable opportunities for financial applications. The utilization of ML and DM methods allow the automatic and instant extraction of important information from large data quantities of diversified sources that can be used for different applications such as bankruptcy prediction, stock market forecasting, credit scoring or fraud detection. Moreover, the generalized utilization of social media services provide large data sets about several aspects of persons lives. This recent data source supplies novel information that can complement and enhance existing financial applications.

2.4.2 *Machine Learning and Data Mining*

At the present time, large amounts of data are continually generated by different sources (e.g., social media, mobile phones, sensors) at diverse velocities and frequencies. For instance, the volume of business data duplicates every 1.2 years according to Chen and Zhang (2014). The high volume, variety and velocity data is commonly designated as Big Data (Laney, 2001). Moreover, the rapid improvement in storage and processing capabilities permits an effective exploration of the vast proliferation of structured and unstructured data. ML and DM tools enable the discovery of valuable information from large amounts of data that can be important to decision making.

Finance is particularly prolific in big data. For example, New York Stock Exchange (NYSE) Euronext was already processing two terabytes of data daily in 2013⁴. The analysis of extensive quantities of financial data may provide competitive advantages for diverse financial operations. In an initial literature review about intelligent data analysis systems in finance, we found that research has been mainly focused in three

⁴ <http://www.ibmbigdatahub.com/sites/default/files/document/NYSE-Euronext-IMC14787USEN.PDF> [Accessed on 2 January 2017]

financial applications: bankruptcy prediction, credit scoring and stock market prediction.

The applications of bankruptcy prediction try to forecast the chances of bankruptcy for companies or individuals. The majority of these applications use a binary classification by labeling each instance as "bankruptcy" / "no bankruptcy" (Lin et al., 2012). Nevertheless, there are other bankruptcy representations using a greater number of classes, proportional to the risk of failure.

The first methods applied to predict bankruptcy were conventional statistical techniques. ML techniques started to be used in this topic in the 90's (Lin et al., 2012). The nonlinear patterns found in bankruptcy prediction problems recommend the utilization of some ML methods rather than conventional statistical techniques (Kumar and Ravi, 2007; Lin et al., 2012). For instance, NN are considered the most applied ML method for bankruptcy prediction (Kumar and Ravi, 2007; Lin et al., 2012).

Hybrid and ensemble models have been increasingly applied in recent years (Kumar and Ravi, 2007; Lin et al., 2012). Both ensemble and hybrid methods may benefit from complementary characteristics of different individual methods. Therefore, these methods may present better results than their individual components (Kumar and Ravi, 2007; Lin et al., 2012). Hybrid models are also more utilized than ensembles for bankruptcy prediction because they require less training and testing and a lower diversity of classifiers (Lin et al., 2012).

Credit scoring applications estimate the creditworthiness of their target entity. For instance, a financial institution may use credit scoring systems to evaluate the risks of conceding a loan to a specific person. The produced score or classification reflects the probability of commitment fulfillment and influences the credit approval decision and the terms of the loan (e.g., interest rate).

The first credit scoring models were based in some statistical techniques such as Linear Discriminant Analysis, Logistic Regression Analysis and Multivariate Adaptive Regression Splines. However, some assumptions of these techniques are frequently violated in financial data sets (e.g., multivariate normality assumptions for independent

variables), which makes them inadequate for these situations (Huang et al., 2004). ML methods such as NN, Decision Trees or SVM, do not assume certain distributions of data and automatically extract knowledge from training samples (Wang et al., 2011). They presented better performance than conventional statistical techniques, especially for classification of nonlinear patterns (Huang et al., 2004).

Most credit scoring models use data related with the information filled in credit application forms and customer's credit information provided by credit agencies. There is also a vast amount of information detained by the organization about previous customers that could be useful for a credit rating application (Thomas, 2000). However, these databases may not be a representative sample of the total population of candidates because they may not incorporate refused candidates. The features used in these models are diverse and may be related to the financial structure, capacity to pay debt, manageability and operational profitability (Wang et al., 2011).

Stock market applications intend to support investors in their investment decisions or even automatically identify and execute trading strategies. These algorithms may predict diverse stock market variables to support trading decisions. Some of these variables are:

- returns: measure of the relative change in a security value.
- volatility: latent measure of total risk associated with a given investment.
- trading volume: number of shares traded.

This information is useful for investors to value their assets, manage their risk exposure or measure liquidity.

There are two main approaches used by investors to create portfolios of securities: fundamental and technical analysis. These approaches are based on distinct assumptions, methods and objectives. The fundamental analysis implies a detailed study of the company in order to assess its intrinsic value. It assumes that the stock price moves towards its intrinsic value. Therefore, it is possible to identify opportunities by detect-

ing differences between current stock prices and the respective value of the company. For instance, the investor may buy the stock when the estimated value is higher than its current market price. The estimation of the value of the company involves the analysis of quantitative information included in financial statements and also the assessment of qualitative factors such as management, industry, patents and brand name. Some common financial measures applied in this analysis are financial ratios such as: the Price-to-Book Ratio, Price-to-Earnings Ratio, Dividend Yield, Debt to Equity Ratio or Returns on Equity.

Technical analysis applies statistics on past market activity, prices and volume. This approach tries to identify patterns and indicators that may suggest some future behavior. It assumes that markets are not efficient and that history repeats. Rather than having solid information about the target company, technical analysts apply technical indicators (e.g., moving average, exponential moving average, moving average convergence/divergence, relative strength index) and try to identify specific patterns (e.g., cup and handle, head and shoulders) on stock charts. It is mainly used to predict the stock behavior for short periods of time while fundamental analysis is focused on longer periods.

Features derived from these analysis approaches are usual inputs for DM and ML models to predict stock market variables. For instance, a significant part uses data related to stock prices (e.g., high price, low price, open price, close price) or measures about fundamental factors (e.g., revenues, earnings per share, capital investment, debt, and market share) (Zhang and Zhou, 2004). The average number of input variables varies between four and ten (Atsalakis and Valavanis, 2009). The stock market prediction is extremely difficult because it is heavily influenced by events of various kinds. Economic, political, international and even natural factors can be crucial in forecasting (Enke and Thawornwong, 2005). Investors sentiment can also play a role on their investment decisions since individuals decision making processes are affected by their emotions (Peterson, 2007). The qualitative information is highly subjective and very difficult to measure.

Most of the initial regression models applied for stock market prediction relied on simple linear regression assumptions (Enke and Thawornwong, 2005). However, the relationship between some stock market variables and financial and economic data may be nonlinear. Therefore, the application of more flexible learning ML methods became more popular in this topic. For instance, NN are considered the most used method for stock market forecasting (Atsalakis and Valavanis, 2009; Yoo et al., 2005; Zhang and Zhou, 2004). Besides its ability to learn the nonlinear relationships between the variables, NN is tolerant to noisy and incomplete data representation. SVM is a widely applied alternative method that is considered less prone to overfitting than NN and obtains the global minimum solution.

Web and social media data have been increasingly applied for stock market prediction because they permit the instant inclusion of diverse types of influential information. For instance, the utilization of investor sentiment and attention indicators extracted from Web data sources have been tested in the forecasting of the stock market behavior (e.g., Antweiler and Frank, 2004; Bollen et al., 2011; Da et al., 2011).

2.4.3 *Sentiment and Attention Indicators*

The standard finance model defends that investors act as rational agents and all existing information is reflected immediately in security prices. Unsentimental investors coerce stock prices to match their present value of expected future cash flows. However, there is a strand of the finance literature (behavioral finance) that argues that sentiment may affect financial prices (Shiller, 2003). Indeed, there are some events that do not seem to be properly explained by the classical model such as the Great Crash of 1929, the Black Monday crash of October 1987, and the 1990s Dot.com bubble (Baker and Wurgler, 2007).

Some recent literature argues that sentiment may affect prices. Long et al. (1990) consider that there are two types of investors with different behavior and exposure to

Chapter 2. Background

sentiment. Noisy investors may trade on non-fundamental information while rational traders act unemotionally and behave as arbitrageurs. Nevertheless, rational investors may not always force prices to fundamental values. Diverse studies identify various limits to arbitrage. Shleifer and Vishny (1997) refer that in some extreme circumstances, specialized professional arbitrageurs may not correct security prices effectively for various reasons, such as capital limitations or agency costs. Long et al. (1990) argue that the unpredictability of irrational traders creates an additional risk on prices that prevents rational investors from vigorously betting against them. In addition, some institutional investors are forbidden by their charters from shorting (Stambaugh et al., 2012). Therefore, security prices may diverge from fundamental values because arbitrage may fail to eliminate mispricing caused by investor sentiment. If it happens, future prices may be predictable and sentiment indicators may have predictive information.

Moreover, some literature defend that investors attention may affect asset prices and dynamics (e.g., Merton, 1987; Hirshleifer and Teoh, 2003). Information processing of the all available stock market information is impracticable because human attention is a scarce cognitive resource. Therefore, investors need to choose the information and may neglect important aspects of their economic environments (Hirshleifer and Teoh, 2003). Thus, important information may not be properly reflected by prices due to limited investor attention. Furthermore, the temporal variation of the investor attention about a company may affect its prices. For instance, if new public information about a company captures more attention than previous releases, then its investor base can increase and its stock price can rise (Merton, 1987).

As shown in Table 2, there is a large list of papers assessing the predictive value of sentiment and attention for stock market behavior that can be distinguished in terms of several dimensions. The sentiment and attention indicators can be created using distinct sources (column *Source*), SA methods (*Meth.*) and combination methods used to merge distinct sources (*Comb.*). The financial analysis assumes a periodicity (*Per.*) of the applied variables (e.g., daily, monthly), type of stock (*Stocks*, e.g., individual or portfolios), methods (*Meth.*) used to model or predict (e.g., MR) and data (*Data*) used to

fit the models (e.g., four months). Some of the most recent studies (after 2011), perform a prediction that is characterized by its data (*Data*) period (e.g., nineteen days) and the statistical tests used to verify the statistical (*St.*) significance of the sentiment and attention based predictions when compared to baseline models. None of the related works attempts to predict survey sentiment indices (*Sur.*). In the next few paragraphs, we detail some of these dimensions.

Different proxies for investor sentiment have been applied and most of them are not related to social media. Some papers use indirect measures such as economic and financial variables to infer the emotional state of investors. For instance, the Baker and Wurgler monthly index (Baker and Wurgler, 2006) was adopted in several studies (Baker and Wurgler, 2007; Kurov, 2010; Yu and Yuan, 2011; Chung et al., 2012; Hribar and McNnis, 2012; Stambaugh et al., 2012; Corredor et al., 2013). This index corresponds to the first principal component of six different proxies for sentiment: the closed-end fund discount, NYSE share turnover, the number and average first-day returns on Initial Public Offerings, the equity share in new issues, and the dividend premium.

Other papers applied direct sentiment measures derived from surveys of investors' feelings about the stock market. For instance, AAI provides weekly values of the votes of their members to a poll questioning their sentiment (bullish, bearish, neutral) on the stock market for the next six months. Research works such as (Fisher and Statman, 2000; Brown and Cliff, 2004; Kurov, 2008; Verma and Verma, 2008; Verma and Soydemir, 2009) used AAI index as a sentiment measure. Also, UMCS is a monthly sentiment index constructed from a consumer confidence survey answered by a random group of five hundred continental US households. Despite being considered less related to investor sentiment, UMCS index has been applied in a considerable number of stock market studies (e.g., Fisher and Statman, 2000; Lemmon and Portniaguina, 2006; Qiu and Welch, 2006; Ho and Hung, 2009; Schmeling, 2009; Zouaoui et al., 2011; Chung et al., 2012; Stambaugh et al., 2012). Moreover, Sentix (www.sentix.de) creates sentiment indices for various stock markets (e.g., US, Japan, Germany, Euro zone)

Chapter 2. Background

Table 2.: Summary of literature about the analysis of the predictive value of sentiment and attention for stock market behavior

Study	Sentiment			Attent.	Financial Analysis			Prediction			
	Source ^a	Meth. ^b	Comb. ^c		Source ^a	Per. ^d	Stocks ^e	Meth. ^f	Data ^g	Data ^g	St. ^h
(Solt and Statman, 1988)	S				w	Ix	MR	22y			
(Lee et al., 1991)	F				m	Pf	MR	20y			
(Neal and Wheatley, 1998)	F				m,q,a	Pf	MR	60y			
(Fisher and Statman, 2000)	S				m	Ix,Pf	MR	13y			
(Tumarkin and Whitelaw, 2001)	MB			MB	d	I	VAR	11m			
(Lee et al., 2002)	S				w	Ix	GARCH	22y			
(Antweiler and Frank, 2004)	MB	ML		MB	d	I	MR	1y			
(Brown and Cliff, 2004)	F,S		KF,Pca		m,w	Pf	VAR	33y			
(Brown and Cliff, 2005)	S				m	Pf	MR	19y			
(Das et al., 2005)	MB,N	ML		MB,N	d	I	MR	7m			
(Baker and Wurgler, 2006)	F		Pca		m	Pf	MR	38y			
(Qiu and Welch, 2006)	F,S				m,q	Pf	MR	38y			
(Schmeling, 2007)	S				w	Ix	MR	4y			
(Das and Chen, 2007)	MB	ML		MB	d	Ix,I	MR	2m			
(Tetlock, 2007)	N	GL			d	Am,Ix,Pf	VAR	15y			
(Ho and Hung, 2009)	S		Pca		m	I	MR	41y			
(Schmeling, 2009)	S				m	Am,Pf	MR	21y			
(Kurov, 2010)	F,S		Pca		d	Ix,I	MR	14y			
(Yu and Yuan, 2011)	F		Pca		m	Am	MR	42y			
(Bollen et al., 2011)	M	GL			d	Ix	NN	11m	19d		
(Deng et al., 2011)	N	GL		N	d	I	2ML,RW	32m	2y		
(Groß-Klußmann and Hautsch, 2011)	N	P			i	I	VAR	18m			
(Mao et al., 2011)	G,M,N,S	FL,K			d,w	Ix	MR	15m	30d,20w		
(Oh and Sheng, 2011)	M	ML		M	d	I	8ML	4m	10d		
(Sabherwal et al., 2011)	MB	ML		MB	d,i	I	MR	13m			
(Sheu and Wei, 2011)	F				d	Am	MR,TR	4y	59d		
(Zhang et al., 2011b)	M	K			d	Ix	Cor	7m			
(Baker et al., 2012)	F		Pca		m	Am,Pf	MR	25y			
(Schumaker et al., 2012)	N	GL			i	I	SVM	23d	23d		
(Stambaugh et al., 2012)	F,S		Pca		m	Pf	MR	42y			
(Chen and Lazer, 2013)	M	GL			d	Am	MR,TR	97d	25-33d		
(Corredor et al., 2013)	F,S		Pca		m	Pf	MR	18y			
(Garcia, 2013)	N	FL			d	Ix,Pf	MR	100y			
(Hagenau et al., 2013)	N	ML			d	I	TR	14y	12y		
(Smailović et al., 2013)	M	ML			d	I	GC	10m			
(Yu et al., 2013)	B,M,MB,N	ML		B,M,MB,N	d	I	MR	3m			
(Sprenger et al., 2014)	M	ML		M	d	I	MR	6m			
(Al Nasser et al., 2015)	M	ML			d	Ix	TR	13m	1y	ST	
(Nguyen et al., 2015)	MB	ML,GL		MB	d	I	SVM	13m	78d		
This thesis	M,S	MFL	KF	M	d,w	Ix,Pf	5ML,RW	35m	350-439d	DM	2S
									91-93w		

^a Sentiment and attention sources: B – blogs, F – financial data, G – Google searches, M – microblogs, MB – message boards, N – news, S – surveys

^b Sentiment analysis method: FL – financial lexicon, GL – generic lexicon, K – keywords, ML – supervised machine learning, MFL – microblog financial lexicon, P – sentiment analysis product

^c Combination method: KF - kalman filter, Pca - principal component analysis

^d Periodicities: a – annual, d – daily, i – intraday, m – monthly, q – quarterly, w – weekly

^e Stocks: Am – aggregated market, I – individual stocks, Ix - indices, Pf - Portfolios

^f Financial analysis method: Cor – correlation, GARCH – generalized autoregressive conditional heteroskedasticity, GC – granger causality, MR – multiple linear regression, nML - n machine learning methods, NN - neural networks, RW – rolling windows, SVM – support vector machine, TR – trading rules, VAR – vector auto-regression

^g Data Period: d – days, m – months, w – weeks, y – years

^h Statistical Test for Out of Sample Evaluation: DM – Diebold-Mariano test, ST – Student’s t-test

ⁱ Prediction of Surveys: nS - n survey sentiment indices

from surveys answered by more than 3,500 participants. Some research works used Sentix indices (e.g., Schmeling, 2007; Corredor et al., 2013; Hengelbrock et al., 2013). Another example is the II weekly index based on over a hundred independent market newsletters, with each newsletter being categorized as bullish, bearish or correction. II measures may be more correlated to institutional sentiment than AAI, because many of the newsletters' authors are market professionals (Brown and Cliff, 2004). The II index is widely applied in behavior finance literature (e.g., Solt and Statman, 1988; Fisher and Statman, 2000; Lee et al., 2002; Brown and Cliff, 2004, 2005; Kurov, 2008; Ho and Hung, 2009; Kurov, 2010; Verma and Verma, 2008; Verma and Soydemir, 2009).

Recently, various studies used investor sentiment and attention indicators automatically extracted from textual contents. Diverse data sources were used to create these indicators, such as newspapers (e.g., New York Times, Wall Street Journal) (Tetlock, 2007; Deng et al., 2011; Groß-Klußmann and Hautsch, 2011; Schumaker et al., 2012; Garcia, 2013; Yu et al., 2013; Geva and Zahavi, 2014), message boards (e.g., thelion.com, ragingbull.com) (Tumarkin and Whitelaw, 2001; Antweiler and Frank, 2004; Das et al., 2005; Das and Chen, 2007; Sabherwal et al., 2011; Yu et al., 2013; Li et al., 2014), or microblogs (e.g., Twitter, StockTwits) (Smailović et al., 2014; Yu et al., 2013; Zhang et al., 2011b; Smailović et al., 2013; Makrehchi et al., 2013; Chen and Lazer, 2013; Bollen et al., 2011; Sprenger et al., 2014; Oh and Sheng, 2011). The indicators extracted from texts (e.g., Twitter) have many advantages when compared with survey sentiment indices. The creation of text based sentiment indicators is faster and cheaper, permits greater periodicities (e.g., daily) and may be targeted to a more restrict set of stocks (e.g., stock market indices or individual stocks). The distribution of sentiment sources applied in this literature review is presented in Table 11.

A significant part of the analyzed papers apply survey or financial data to generate sentiment measures. Therefore, the application of SA methods occurs only in half of these studies. There are two main approaches for the extraction of sentiment indicators from text: supervised and unsupervised. Some studies use supervised ML, such as NB or SVM (Antweiler and Frank, 2004; Hagenau et al., 2013) but it requires labeled

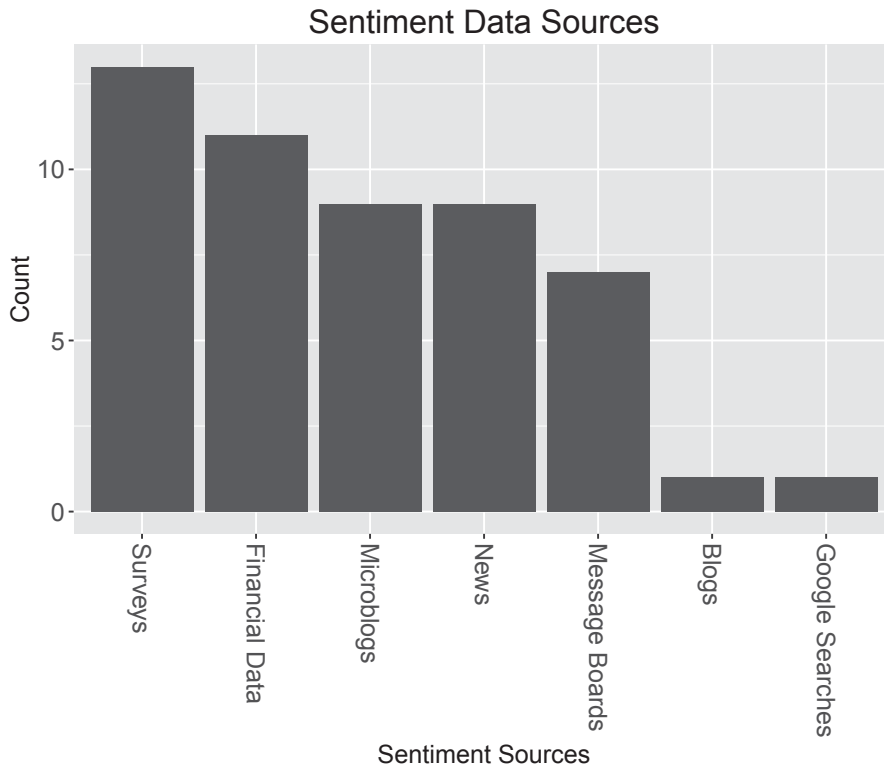


Figure 11.: Distribution of sentiment data sources

training data that is often difficult to obtain. Social media often do not provide classified data and their manual labeling is costly and impractical. Thus, other studies use unsupervised approaches based on lexicons or keywords (Bollen et al., 2011; Mao et al., 2011). Most of the applied lexicons are domain independent (e.g., GI, MPQA, SWN). Only two studies use the financial lexicon created by (Loughran and McDonald, 2011). Generic domain independent lexicons may be ineffective for assessing the sentiment of stock market messages because some terms may have different sentiment values in distinct contexts. Moreover, the financial lexicon of (Loughran and McDonald, 2011) was created using large text reports and it may not be properly adjusted for short microblogging messages. The application of a large lexicon adapted to microblogging stock market conversations should allow the production of more robust sentiment indicators. Table 12 shows the occurrence of SA approaches.

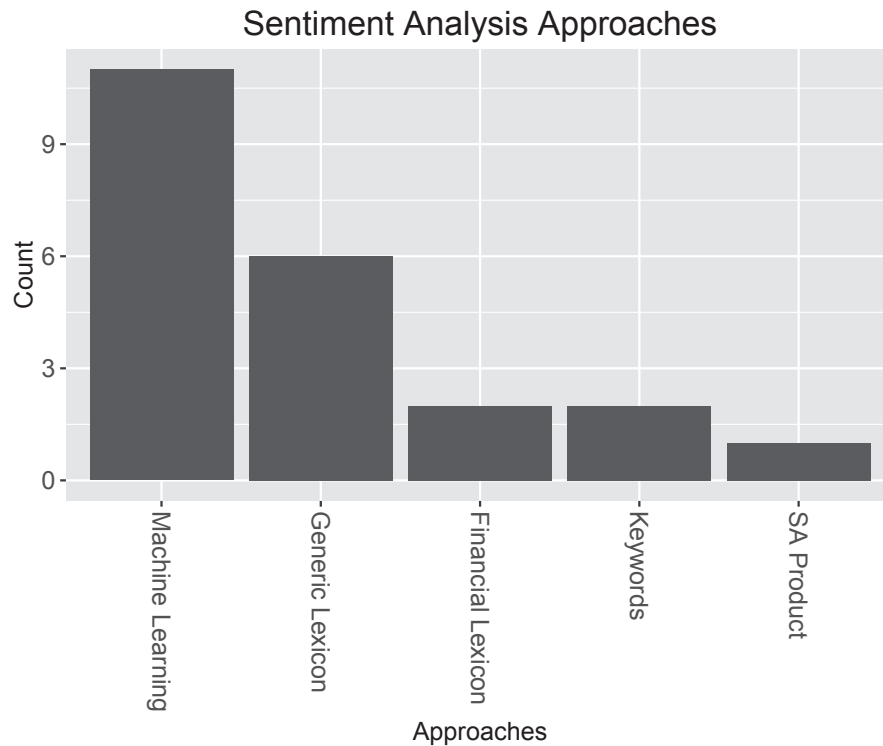


Figure 12.: Distribution of sentiment analysis approaches

Furthermore, the various existing investor sentiment indicators are measured by different approaches (e.g., surveys, social media) and have distinct characteristics (e.g., monthly, weekly or daily frequencies). Therefore, they usually have different values and may contain some noise. Some studies (17.5%) combine different financial measures to create a unique sentiment indicator. For instance, Baker and Wurgler use Principal Component Analysis to extract a sentiment indicator from six financial measures (Baker and Wurgler, 2006). This procedure has been applied in some studies using financial data and surveys and for low frequencies (e.g., monthly, weekly). Only one study combined two different types of sentiment sources. A KF procedure was applied to extract a weekly and a monthly sentiment indicator from financial data and surveys (Brown and Cliff, 2004). To the best of our knowledge, there were no previous attempts in this topic to produce sentiment indicators by combining social media with other sources, particularly for higher frequencies than the weekly.

Chapter 2. Background

The usage of attention proxies is infrequent in this review because only 27.5% of the papers used attention indicators. The number of posts in message boards, the number of news related to the studied stocks and the number of microblogging messages are the usual attention indicators.

As shown in Table 2, the size of the data sets applied in this topic is usually high for studies using sentiment based on financial data, surveys or news but it is low for sentiment extracted from social media data. These latter studies use less than two years of data. The utilization of larger periods of social media data should permit a more robust analysis of the informative content of these data for stock market predictions.

The evaluation of sentiment on portfolios based on some characteristics (e.g., size, book to market, volatility) is frequent on financial data or surveys studies. However, this analysis is very scarce for text based sentiment and for higher periodicities than the weekly. For instance, Tetlock (2007) and Garcia (2013) analyzed the influence of sentiment created from news on some variables based on portfolios formed on size. We did not find any analysis of the influence of sentiment based on social media on portfolios of any type. These studies focus on the prediction of individual stocks, indices or aggregated market. Thus, it would be interesting to assess the impact of sentiment extracted from social media data on different portfolios, particularly for high periodicities (e.g., daily). Figure 13 shows the frequency of predicted stock types.

To evaluate the predictive value of sentiment for stock market variables, the majority of the studies apply MR and VAR methods, which are generalized for survey and financial based sentiment and very frequent for text based sentiment. The usage of more flexible learning ML models, such as SVM (Deng et al., 2011; Nguyen et al., 2015; Schumaker et al., 2012) or NN (Bollen et al., 2011), is more scarce. Also, very few studies compared the predictive accuracy of different ML models for stock market variables (Oh and Sheng, 2011; Deng et al., 2011). The distribution of the applied methods is presented in Figure 14.

Most forecasting studies that use text extracted indicators present limitations in terms of lack of a robust evaluation. For instance, a out of sample evaluation was

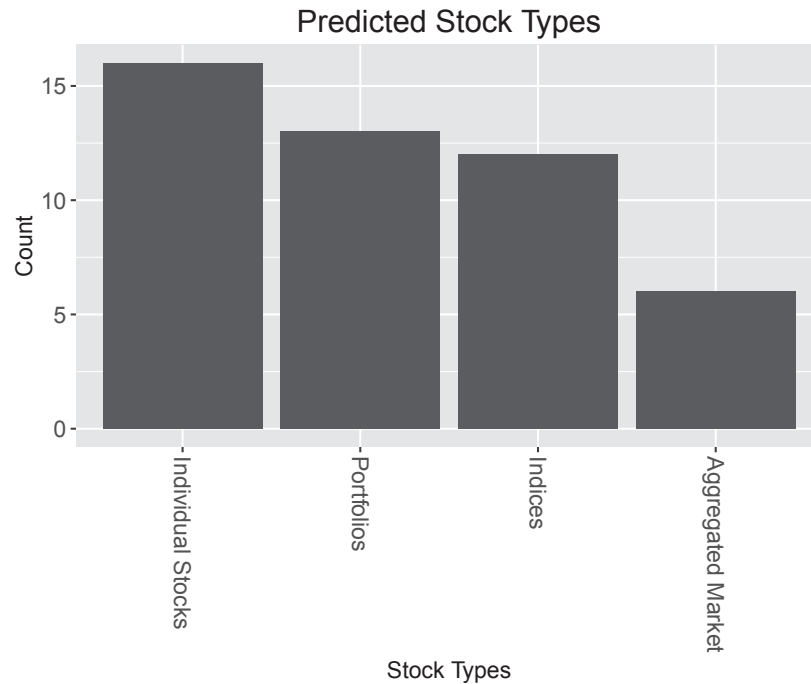


Figure 13.: Distribution of predicted stock types

not used in almost all papers using financial data or surveys and in the majority of the studies using textual contents. Other works applied very short test sets: 10 predictions (Oh and Sheng, 2011); 19 forecasts (Bollen et al., 2011); 20 and 30 forecasts (Mao et al., 2011); 25 and 35 predictions (Chen and Lazer, 2013); and 79 forecasts (Nguyen et al., 2015). Moreover, the utilization of statistical tests to evaluate the predictive accuracy is very limited (Table 2). The application of a statistical test of predictive accuracy to assess the informative value of social media data would produce more solid evaluation results.

The prediction of survey sentiment indices can be very useful for investors. It may permit a valuable anticipation of their values or constitute a cheap alternative measure. Though there are regressions of survey sentiment using contemporaneous values of other sources (e.g., financial measures, other surveys) in some studies (Brown and Cliff, 2004; Schmeling, 2007), we did not find their prediction based on lagged values of other sentiment proxies, particularly social media sentiment.

The results obtained by this research topic are illustrated in Figure 15.

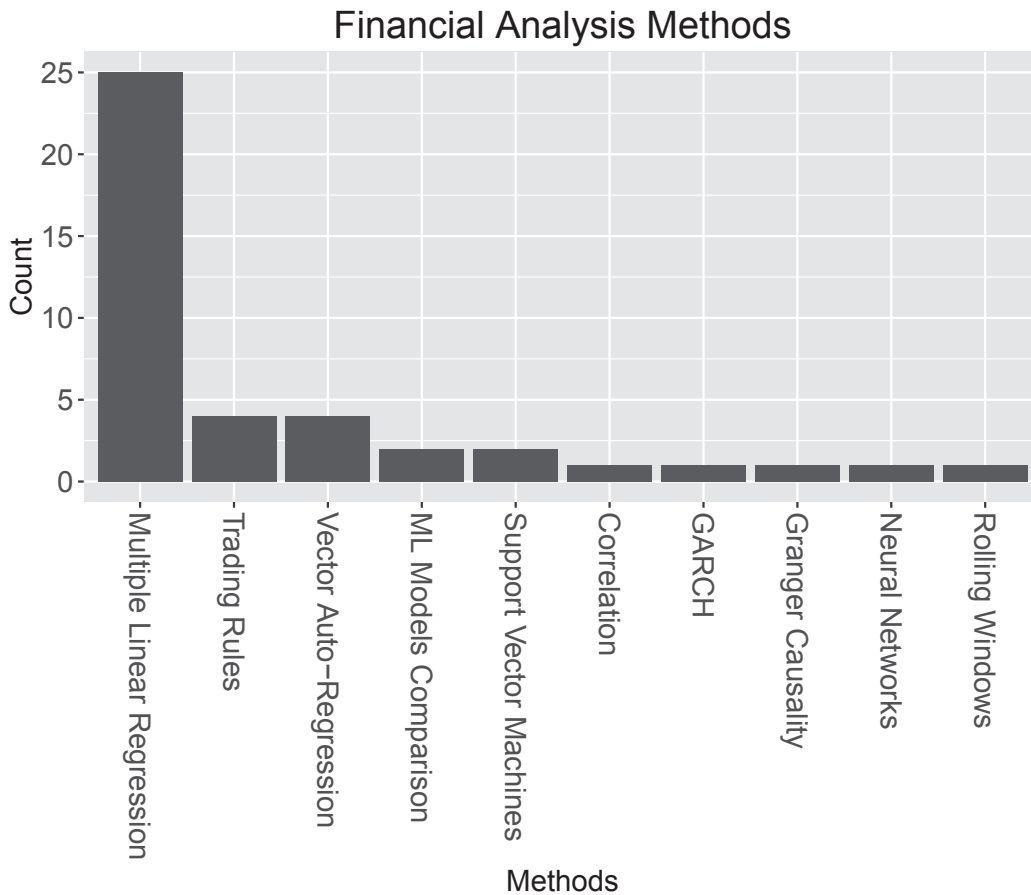


Figure 14.: Distribution of methods applied for financial analysis

Text based sentiment was considered useful to make trading decisions (Schumaker et al., 2012; Al Nasserri et al., 2015) or predict useful stock market variables, such as: daily or intraday values of stock prices (Bollen et al., 2011), price directions (Nguyen et al., 2015), returns (Tetlock, 2007; Sabherwal et al., 2011), volatility (Sabherwal et al., 2011) and trading volume (Antweiler and Frank, 2004). The analysis of the informative value of sentiment extracted from text has been almost exclusively focused on the prediction of daily variables of individual stocks, indices or aggregate market. However, sentiment collected from surveys and financial data has been mainly applied for lower periodicities (e.g., monthly) and portfolios formed based on diverse characteristics (e.g., firm dimension, age, volatility, book-to-market ratio). Sentiment seems to have more effect on returns of some portfolios having extreme values on these charac-

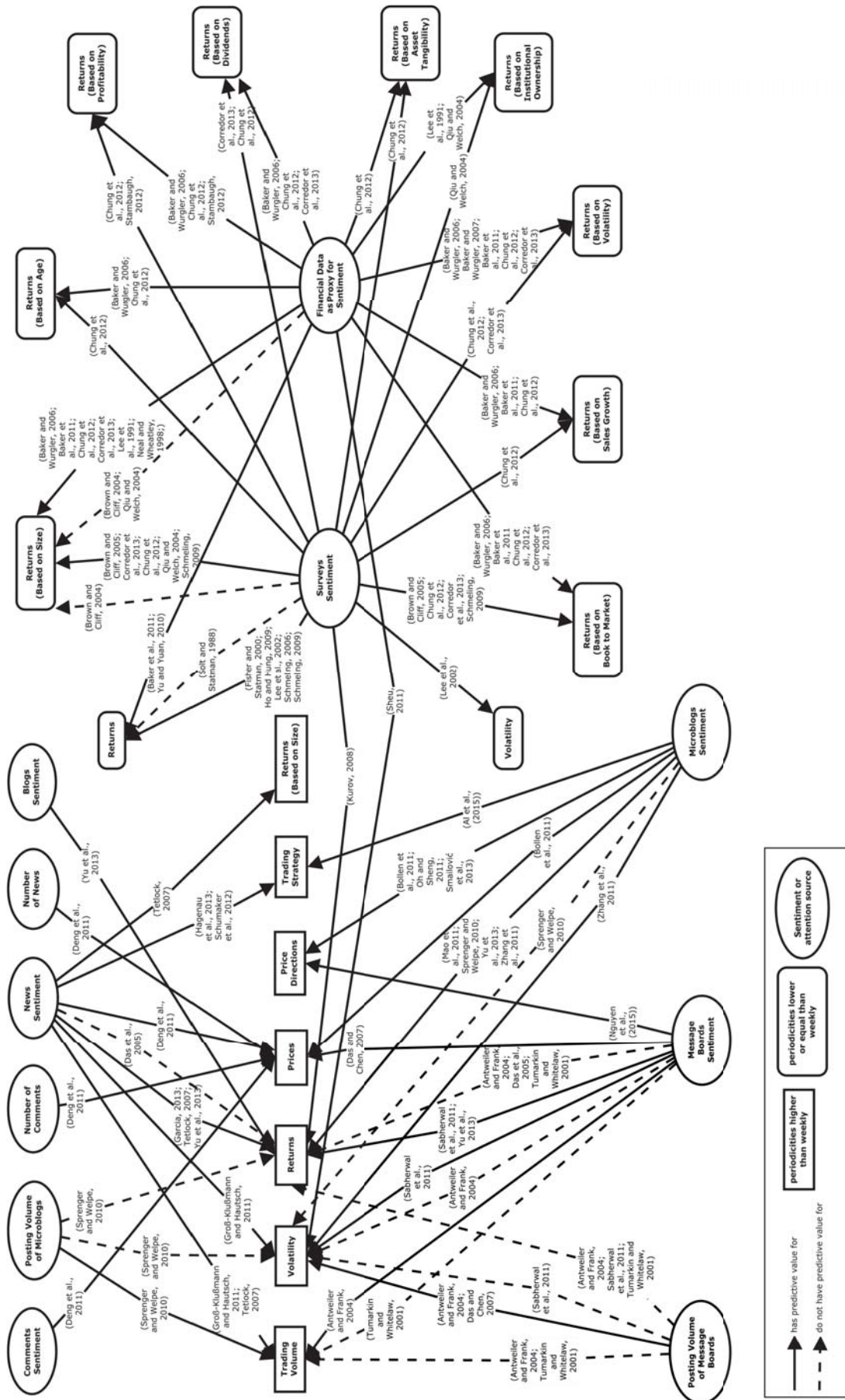


Figure 15.: Results obtained by the literature about the predictive value of sentiment and attention for stock market behavior

Chapter 2. Background

teristics (Neal and Wheatley, 1998; Brown and Cliff, 2005; Baker et al., 2012). To the best of our knowledge, there are no studies using microblogging data to predict variables based on portfolios. Posting volume on social media services (e.g., microblogs and message boards) has also been applied to predict volatility (Antweiler and Frank, 2004; Das and Chen, 2007) or trading volume (Sprenger et al., 2014).

Yet, these findings are not consensual and there are studies that find scarce evidence of the predictive power of sentiment for stock prices or returns (e.g., Tumarkin and Whitelaw, 2001; Antweiler and Frank, 2004; Timmermann, 2008; Brown and Cliff, 2004) and posting volume for volatility (Antweiler and Frank, 2004; Tumarkin and Whitelaw, 2001).

2.5 SUMMARY

This chapter presents the main background knowledge of this thesis. The evaluation of the predictive value of sentiment and attention indicators extracted from social media data for stock market behavior is a multidisciplinary topic involving distinct areas such as Finance, DM, ML, SA, TM and Statistics.

Three main areas are presented in this chapter. Section 2.2 explains the importance of DM and ML for this project and describes the five applied DM methods:

- MR assumes a linear association between the dependent target and various independent variables.
- NN is a system of neurons interconnected by numeric weights calibrated during the learning process.
- SVM use linear models to obtain nonlinear associations by converting the input data into a higher dimensional space using kernel functions.
- RF produces a large number of unpruned decision trees and calculates the final output by aggregating each individual output.

- Ensembles combines the predictions of diverse individual models.

These methods are important to create predictive data-driven models for stock market variables. Each regression method has its own learning capabilities and advantages. MR is more easy to interpret, does not require the tuning of hyperparameters and has less tendency to overfit. Yet, MR is a rather rigid model and it has limited capacity to extract nonlinear associations. All other methods (NN, SVM and RF) are more flexible and suited to deal with nonlinear complex variable relationships. Often, these nonlinear methods lead to complex data-driven models. Functional nonlinear methods, such as NN and SVM, are more prone to overfitting and thus several hyperparameters need to be tuned (e.g., hidden nodes, kernel parameter). In contrast, RF tends to provide good results with its default parameters but for several small or medium sized data sets the computational effort to fit the model is often larger than a single NN or SVM training due to the usage of a large number of trees. When compared with NN, SVM present theoretical advantages, such as the absence of local minima in the model optimization phase. Ensembles frequently obtain better predictive performances than their individual models (Oztekin et al., 2016) and some approaches (e.g., EA) are very simple and do not require a substantial computational effort.

In this project, SA is essential to create solid investor sentiment indicators from social media data. SA is the computational analysis of opinion, sentiment and subjectivity in text (Pang and Lee, 2008). Textual opinionated contents have certain characteristics that distinguishes them from other data. For instance, text is unstructured, sparse and high dimensional which imposes additional difficulties to common DM problems. SA processes these specificities by representing text using diverse features such as term frequencies, term position in the document, opinion term presence, POS, dependency relations or word embeddings. These different text representations can be used as inputs to ML models for supervised SA. However, labeled social media data is often very difficult to obtain. Thus, the utilization of opinion lexicons to perform unsupervised SA is also very common. Diverse publicly available generic lexicons are applied but

Chapter 2. Background

they may not be efficient for stock market contents. For instance, there are several terms and expressions that have different sentiment value in the stock market context. Moreover, stock market conversations have a specific terminology that may not be included in generic lexicons. Therefore, the utilization of lexicons properly adapted to stock market and social media contents can be very important for a fast, light and reliable unsupervised SA in this domain.

The understanding of the basic concepts of financial markets is also important for this project. Financial markets aggregate the activities of various stock market agents by coordinating the transaction of stocks. Companies can raise capital by selling their stocks and investors can create portfolios by trading them. A rigorous analysis of information related to these stock market variables (e.g., returns, trading volume, volatility) can be used to inform investment decisions, as estimate potential risk exposure, or to estimate stock liquidity.

Large volumes of financial data can be used by financial entities to support their decisions. ML and DM methods are instrumental to extract informative patterns from these data. These methods have been used for diverse financial applications such as credit approval, bankruptcy prediction or fraud detection. The exponential spread of Social Media data can also be explored for financial prediction. The analysis of social media contents may provide an accurate assessment of their users' sentiment. Therefore, the application of SA on social media data can be useful to analyze the influence of sentiment on stock market behavior. Social media sentiment indicators may replace or complement traditional sentiment sources (e.g., surveys) with some advantages. The extraction of sentiment from social media data is faster, cheaper and permits a more flexible sentiment creation (e.g., for different periodicities or particular stocks).

This chapter also presented a comprehensive review of the literature about the evaluation of the predictive value of sentiment and attention for stock market behavior. Several studies found informative content on sentiment or attention for the prediction

of stock market variables. However, these findings are not consensual because many of these results are not supported by other studies.

In this topic, we could identify some research opportunities. For instance, many investor sentiment indicators are created by lexicon based methods. However, most of them use generic lexicons that may be inadequate for stock market and social media contents. Moreover, ML models need labeled data sets that are very difficult to obtain and could be computationally expensive. Therefore, the utilization of a large lexicon properly adapted to microblogging stock market messages could provide robust investor sentiment indicators in a fast and easy unsupervised SA process.

To the best of our knowledge, there were no previous studies combining social media with other sources in order to produce an unique daily sentiment indicator. The aggregation of sentiment sources with distinct frequencies could create a less noisy and more representative investor sentiment measure.

The assessment of the impact of sentiment on portfolios is scarcely performed for social media based indicators and for higher periodicities than the weekly. The utilization of social media data to predict short term returns of portfolios could address the analysis of the short term predictive value of sentiment on different series of financial returns.

Most studies use MR or VAR models to assess the predictive value of sentiment or attention for stock market behavior. The utilization of more flexible learning ML methods is limited. Also, very few studies apply and compare more than one model for this purpose. Thus, the application of diverse ML methods could produce more accurate prediction results.

Studies using social media data to predict stock market behavior usually apply short data sets (e.g., less than two years of data). Also, the majority of the studies do not perform out of sample evaluation or use short test sets. Furthermore, the utilization of statistical tests of predictive accuracy is very scarce. Thus, the utilization of a statistical test to assess the predictive accuracy on large test sets would produce more robust

Chapter 2. Background

evaluation results about the predictive value of social media data for stock market variables.

The forecasting of survey sentiment values may allow a valuable anticipation of their values or provide an acceptable alternative. We did not find studies predicting survey sentiment by applying lagged values of other sentiment sources such as social media sentiment.

This work attempts to address the identified research opportunities (last row of Table 2) in order to perform a more robust analysis of the informative content of social media sentiment and attention for stock market prediction.

Part II

MAIN BODY

CREATION OF A SPECIALIZED STOCK MARKET LEXICON

3.1 INTRODUCTION

Recently, social media (e.g., Twitter, Facebook, message boards) has enabled a burst of unstructured opinion content that is potentially valuable for diverse decision-making processes (Montoyo et al., 2012). Due to the volume and velocity properties of social media data, human analysis is impracticable and thus SA is used to automatically mine large amounts of opinionated contents in order to summarize the opinions (Hu and Liu, 2004). Several SA approaches apply common supervised classifiers such as SVM (Kiritchenko et al., 2014), NB (Saif et al., 2012) or ensembles (da Silva et al., 2014; Fersini et al., 2014). However, it is very difficult to obtain labeled data and ML models often require very high computational capacities (e.g., processing power, memory) and excessive training time. Thus, it may be impracticable for nearly real-time SA in large data sets such as the creation of investor sentiment indicators from millions of microblogging stock market messages. As such, the utilization of sentiment lexicons allows a fast and lighter alternative unsupervised classification of text, relieving the need for arduous manual labeling of text and expensive computer resources. Moreover, sentiment lexicons permit the creation of important features for supervised SA (Rosenthal et al., 2015b). The sentiment lexicon is a list of words with a sentiment value (e.g., positive, negative) and it is considered a key element for SA (Feldman, 2013). For

example, the sentence “this car is great” can be easily detected as positive if the lexicon has a term “great” with a positive value.

SA is being increasingly used to predict stock market variables (Antweiler and Frank, 2004; Schumaker et al., 2012; Schumaker and Chen, 2009; Yu et al., 2013). In particular, microblogging data are a useful source for supporting stock market decisions (Bollen et al., 2011). Users post very frequently and data are readily available at low cost, allowing real-time assessment that can be exploited during the trading day. However, there has been little effort in producing lexicons adapted to the financial domain and microblogs. A financial lexicon was manually built by Loughran and McDonald (2011) using text documents extracted from the U.S. Securities and Exchange Commission portal from 1994 to 2008. Mao et al. (2014) proposed a procedure to automatically construct a Chinese financial lexicon by exploring a large news corpus classified as positive or negative according to the contemporaneous stock returns. Yet, these lexicons did not consider microblog messages, which is often informal and has character constraints. Furthermore, adopting a manual approach (e.g., Loughran and McDonald, 2011) is not feasible in practical terms given the huge effort required to label the large volumes of microblog texts. Moreover, the existing domain independent lexicons (e.g., Stone et al., 1966; Wilson et al., 2005; Baccianella et al., 2010) may be ineffective for stock market contents. For instance, the word “explosive” is negative in most contexts but it may be positive in financial messages (e.g., “explosive rise”).

In this chapter, we present a novel automated approach for the acquisition of microblog stock market lexicons. The main contributions are:

- 1) The adaptation of three statistical measures (e.g., PMI) and creation of two new complementary statistics. These measures are applied in labeled messages of the StockTwits microblogging service to calculate a stock market sentiment score. To address negation more efficiently, sentiment scores are also created for affirmative and negated contexts.

3.2. Automatic Procedure for the Creation of a Stock Market Lexicon

- ii) The comparison of the resulting stock market lexicons created using the StockTwits test data with six large popular lexical resources: GI (Stone et al., 1966), Opinion Lexicon (OL) (Hu and Liu, 2004), Macquarie Semantic Orientation Lexicon (MSOL) (Mohammad et al., 2009), MPQA subjectivity lexicon (Wilson et al., 2005), SWN (Baccianella et al., 2010) and Financial Sentiment Dictionaries (FIN) (Loughran and McDonald, 2011).
- iii) The assessment of the utility of sentiment indicators produced with a created and a baseline lexicons using a different microblog data source (Twitter). The new Twitter sentiment indicators are correlated with two traditional survey sentiment indicators: II and AAI.

This chapter is structured as follows. Section 3.2 describes the proposed procedure to automatically create a stock market lexicon. Section 3.3 presents the evaluation scheme that was used to assess the informative content of the various created lexicons and six baseline lexicons. Section 3.4 explains the production of the investor sentiment indicators that were used in the calculation of the correlation between microblogging and survey indicators. Section 3.5 describes the experiments and analyzes the obtained results. Finally, this work is summarized and conclusions are drawn in Section 3.6.

3.2 AUTOMATIC PROCEDURE FOR THE CREATION OF A STOCK MARKET LEXICON

This project presents a novel automated approach for the creation of microblog stock market lexicons. Three adaptations (e.g., PMI) and two new proposed statistics were applied in a large data set of classified data provided by a microblogging platform exclusively dedicated to stock markets (stocktwits.com). StockTwits is a microblogging service exclusively dedicated to stock market conversations (stocktwits.com) that has currently more than 300,000 users. StockTwits users can label their own text messages as “bullish” (optimistic opinion) or “bearish” (pessimistic view). These lexicons are created using StockTwits labeled messages from June 2, 2010 to March 31, 2013, in a total

of 350,000 posts¹. We note that such dimension is significantly higher when compared with the majority of works on this topic.

Additionally, sentiment scores were created for affirmative and negated contexts in order to address negation properly. Lexical items may have distinct sentiment effects in negated segments because some words may have their sentiment inverted while others may have their sentiment strength decreased. So, the utilization of distinct sentiment values for affirmative and negated contexts may improve sentiment classification. To the best of our knowledge, this study presents the first stock market lexicons with two different context scores for each entry.

This work applies two different validation approaches. To build a single large lexicon and evaluate its performance, we adopt a holdout split method, where the first 75% StockTwits classified messages are used to create lexicons (training set) and the remaining (most recent) 25% posts are used for evaluation purposes (test set). To perform a robust comparison of the distinct lexicon creation methods, we adopt a realistic rolling window method (Moro et al., 2014), where the labeled data are split into 20 equally sized parts ordered by time. The first 2/3 messages (training set) of each data window is utilized to create lexicons and the last 1/3 (test set) are applied in the evaluation. After performing the data pre-processing tasks, we selected all items having a minimum number of occurrences in the training set (O_{min}). The removal of non-frequent items is a usual preprocessing task in the creation of lexicons (e.g., Qiu et al., 2011; Loughran and McDonald, 2011; Kiritchenko et al., 2014; Mao et al., 2014). This operation permits the elimination of many orthographic errors. Also, some statistical measures (e.g., PMI) are unsatisfactory estimators of association for infrequent terms (Kiritchenko et al., 2014). Since the dimension of training data sets for each evaluation procedure is very different, we defined a different minimum number of occurrences (O_{min}) for each evaluation scheme.

The usage of different combinations of statistical measures on training data generates several lexicons. First, we produce three lexicons by applying adaptations of three

¹ The total number of StockTwits messages over the same period is nearly 6 millions.

known statistical measures (TFIDF, IG and PMI) to calculate the sentiment score of each selected item. Then, we create three more versions of previous lexicons (i.e., a total of 12 lexicon versions) by utilizing two new complementary statistics ($P_{\text{days}}(l)$ and $M_{\text{assoc}}(l)$). In order to refine the sentiment score, the value obtained by the latter measures is multiplied by the score produced by each adapted measure. Additionally, we tested the calculation of sentiment scores for affirmative and negated contexts. The previously described procedure is used for separate affirmative and negated training data.

3.2.1 Data Pre-Processing

In order to prepare the microblogging data for the lexicon creation, we performed various pre-processing tasks using the R tool (R Core Team, 2013):

- substitute all cashtags (i.e., “\$” character and the respective ticker (e.g., \$AAPL)) by a unique term, thus avoiding cashtags to gain a sentiment value related with a particular time period;
- replace numbers by a single tag, since the whole set of distinct numbers is too vast;
- for privacy reasons, all mentions and URL addresses were normalized to “@user” and “URL”, respectively;
- exclude messages composed only by cashtags, URL links, mentions or punctuation (7,176 messages were removed).

Next, we adopted the Stanford CoreNLP tool (Toutanova et al., 2003) to execute common NLP operations, such as tokenization, POS tagging and lemmatization.

The holdout split method uses a large training set with 250,000 posts and thousands of distinct terms. Thus, for this evaluation scheme we included only terms with more

than $O_{min} = 10$ occurrences in the training data set and excluded all punctuation, resulting in approximately 7,000 unigrams and 27,000 bigrams analyzed. The rolling window method creates lexicons in each one of the 20 data partitions. In this validation scheme, the training set of the partitions is much smaller (about 11,100 messages) and thus we adopted a lower minimum number of occurrences, with $O_{min} = 4$ in each training set. Also, we eliminated all punctuation.

3.2.2 Statistical Measures

In this work, we adapt three popular statistical measures and propose two new complementary ones. The former measures were applied to determine the information value of lexical items and thus allow us to discriminate them between “bullish” or “bearish”:

1. TFIDF – often used for textual data representation (e.g., Turney and Littman, 2003) and that is calculated as:

$$tf(l, d) = \frac{n_{d,l}}{n_D} \quad (20)$$

$$idf(l) = \log \frac{N_d}{N_l + 1} \quad (21)$$

$$tf-idf(l, d) = tf(l, d) \times idf(l) \quad (22)$$

where l is a lexical entry, d is a particular document, $n_{d,l}$ is the number of occurrences of l in document d , n_D is the number of lexical items in document d , N_d is the number of documents and N_l is the number of documents containing l . We first created two documents composed by all messages of each class (d_1 – bullish and d_2 – bearish). Then, we executed the *tfidf* function of the *textir* R package to compute $tf-idf(l, d)$. To provide a single value that reflects the tendency to a sentiment class, we calculated the sentiment value S_{TF-IDF} as:

$$S_{TF-IDF}(l) = \frac{tf-idf(l, d_1) - tf-idf(l, d_2)}{tf-idf(l, d_1) + tf-idf(l, d_2)} \quad (23)$$

The final sentiment class depends on the $S_{TF-IDF}(l)$ value: “bullish” if positive, “bearish” if negative or “neutral” if zero.

2. IG – commonly used to assess the information value of an attribute (e.g., Lewis, 1992; Zheng et al., 2004) and that is computed as:

$$IG(l, c) = \sum_{d \in \{c, \bar{c}\}} \sum_{w \in \{l, \bar{l}\}} p(w, d) \log \frac{p(w, d)}{p(w) \times p(d)} \quad (24)$$

where c refers to a category (bullish or bearish), \bar{c} means the non-membership in category c and \bar{l} refers to the absence of l . Since there are only 2 categories, \bar{c} of each class corresponds to c of the other class and IG values are equal for both categories. Hence, we propose the slight adaptation:

$$\begin{aligned} IG_a(l) = & p(l, bl) \log \frac{p(l, bl)}{p(l) \times p(bl)} \\ & + p(\bar{l}, br) \log \frac{p(\bar{l}, br)}{p(\bar{l}) \times p(br)} - p(\bar{l}, bl) \log \frac{p(\bar{l}, bl)}{p(\bar{l}) \times p(bl)} \\ & - p(l, br) \log \frac{p(l, br)}{p(l) \times p(br)} \quad (25) \end{aligned}$$

where bl refers to bullish class and br corresponds to bearish category. In this calculation, instead of summing the values of mutual information of all tuples, we add those referring to tuples correlated to bullish class, (l, bl) and (\bar{l}, br) , and subtract values corresponding to tuples associated to bearish class, (l, br) and (\bar{l}, bl) . Thus, a positive value indicates a bullish orientation and a negative value means a bearish item. Since very frequent words tend to present very high IG values, we prevent this effect by computing the final sentiment score as:

$$S_{IG}(l) = \frac{IG_a(l)}{n_l} \quad (26)$$

where n_l is the number of times that term l appears in all texts.

3. PMI – a popular statistic in the development of lexicons (e.g., Turney and Littman, 2003; Kiritchenko et al., 2014):

$$PMI(x, y) = \log_2 \frac{p(x, y)}{p(x)p(y)} \quad (27)$$

where x and y are words or sets of words, $p(x, y)$ is the probability that they co-occur, and $p(x)$ and $p(y)$ are the probabilities of occurring x and y in the corpus, respectively. PMI will be largely positive if x and y are strongly associated, highly negative if they are complementary and near zero if there is no significant relationship between them. We adapt the sentiment score to include both positive and negative PMI values:

$$S_{PMI}(l) = PMI(l, \textit{bullish}) - PMI(l, \textit{bearish}) \quad (28)$$

where l is a lexical item, *bullish* refers to all bullish messages and *bearish* corresponds to all bearish messages. The sentiment score signal reflects the sentiment orientation.

The three statistical measures were computed to both unigrams (individual words) and bigrams (two sequential terms). We produced one lexicon for each measure, which includes unigrams and bigrams that present a better sentiment score than their constituent terms.

Two novel complementary metrics, $P_{\text{days}}(l)$ and $M_{\text{assoc}}(l)$, are proposed to refine the sentiment score produced by each previously described metric ($S_{\text{Tfidf}}(l)$, $S_{\text{IC}}(l)$ or $S_{\text{PMI}}(l)$). They may increase or decrease the sentiment value calculated by each of the three adapted metrics. We apply the complementary metrics by multiplying them with other metrics (e.g., $S_{\text{Tfidf}}(l)$, $S_{\text{IC}}(l)$ and $S_{\text{PMI}}(l)$). For example, the score of item l produced by the combination of $P_{\text{days}}(l)$, $M_{\text{assoc}}(l)$ and $S_{\text{PMI}}(l)$ is: $P_{\text{days}}(l) \times M_{\text{assoc}}(l) \times S_{\text{PMI}}(l)$.

The $P_{\text{days}}(l)$ statistic calculates, for each lexical item, the percentage of days where the majority of messages mentioning it have the same sentiment polarity of the item.

To have a less biased measure favouring the dominant class (i.e., bullish), we multiply the daily number of bearish messages containing the lexical item by the following adjustment value:

$$V_{\text{adj}} = \frac{N_{\text{bull}}}{N_{\text{bear}}} \quad (29)$$

where N_{bull} is the total number of bullish messages in training set and N_{bear} is the total number of bearish messages. We tested the $P_{\text{days}}(l)$ metric to prevent terms appearing in an abnormally high number in few days to have a polluted sentiment score by the predominant opinion in those days. While a low value may indicate the existence of the described situation, a high $P_{\text{days}}(l)$ value means that the lexical item has consistently the same sentiment orientation. Therefore, we expect that this measure may improve sentiment score computation.

Previous measures also do not account for the association of two sets of words: intensifiers (e.g., more, increase, up) and diminishers (e.g., less, decrease, down). Yet, the analysis of these relationships may improve the calculation of sentiment. The presence of diminishers may reverse the sentiment of the following word (e.g., less debt) while intensifiers may reinforce it (e.g., more debt). Thus, previous measures will not be effective in those situations. For instance, the likely presence of “less profit” in a negative message would incorrectly decrease the sentiment score of the positive word “profit” when calculated by former statistical measures. Therefore, we propose the $M_{\text{assoc}}(l)$ metric to address this issue:

$$N_{\text{Int}}(l) = N_{\text{IntBull}}(l) + N_{\text{DimBear}}(l) \quad (30)$$

$$N_{\text{Dim}}(l) = N_{\text{IntBear}}(l) + N_{\text{DimBull}}(l) \quad (31)$$

$$M_{\text{assoc}}(l) = \begin{cases} \frac{N_{\text{Int}}(l)}{N_{\text{Int}}(l) + N_{\text{Dim}}(l) \times \frac{T_{\text{Int}}}{T_{\text{Dim}}}} + 0.5 & \text{if } l \text{ is Bullish} \\ \frac{N_{\text{Dim}}(l)}{N_{\text{Int}}(l) + N_{\text{Dim}}(l) \times \frac{T_{\text{Int}}}{T_{\text{Dim}}}} + 0.5 & \text{if } l \text{ is Bearish} \end{cases} \quad (32)$$

where N denotes the number of occurrences of the lexical item adjoined to: IntBull – intensifier words (e.g., more profit) in bullish messages; IntBear – intensifier words in

Chapter 3. Creation of a Specialized Stock Market Lexicon

bearish messages; $T_{DimBull}$ – diminisher words (e.g., less profit) in bullish messages; and $T_{DimBear}$ – diminisher words in bearish messages. For all analyzed elements, T_{Int} is the sum of $N_{Int}(l)$ and T_{Dim} is the sum of $N_{Dim}(l)$.

The $M_{assoc}(l)$ measure is only used in elements with more than four occurrences adjoined to intensifiers and diminishers. We selected this threshold value because we consider that a lower number would produce many cases of less solid values of association. For instance, it is more likely to happen an excessively high $M_{assoc}(l)$ value for elements with two occurrences (e.g., $N_{Int}(l) = 2$ and $N_{Dim}(l) = 0$ for a bullish term) than with four or more occurrences.

We distinguished the formula for bullish and bearish items because bullish terms shall have higher $N_{Int}(l)$ values and bearish terms shall produce higher $N_{Dim}(l)$ values. Since $M_{assoc}(l) \in [0.5, 1.5]$, a $M_{assoc}(l)$ value close to 1.5 means that these associations are highly concordant to the assigned sentiment polarity and the absolute sentiment score will increase. A low $M_{assoc}(l)$ value indicates the opposite, decreasing the absolute score. The intensifiers and diminishers (Table 3) were manually selected. First, we choose a small set of words (e.g., less, more, very) and then added synonyms found in a thesaurus.

Table 3.: Intensifiers and Diminishers

Intensifiers	Diminishers
accretion, accrual, addendum, addition, augmentation, boost, expansion, gain, increment, more, plus, proliferation, raise, rise, accelerate, add, aggrandize, amplify, augment, enlarge, escalate, expand, extend, hype, multiply, swell, stoke, supersize, up, accumulate, climb, proliferate, soar, uprise, desire, fancy, prefer, enjoy, relish, admire, adore, esteem, hallow, idolize, revere, venerate, worship, appreciate, love, elevated, escalated, heightened, increased, raised, admiring, applauding, appreciative, approbatory, approving, commendatory, complimentary, friendly, good, positive	abatement, decline, decrease, decrement, depletion, diminishment, diminution, fall, lessening, loss, lowering, reduction, shrinkage, diminish, dwindle, lessen, recede, wane, abate, downsize, lower, minify, reduce, subtract, drop, descend, dip, plunge, dive, sink, slide, abhor, abominate, despise, detest, execrate, loathe, deplore, deprecate, disapprove, disdain, disfavor, dislike, hate, decreased, depressed, dropped, receded, under, down, low, adverse, depreciative, depreciatory, derogatory, disapproving, inappreciative, negative, unappreciative, uncomplimentary, unfavorable, unflattering, unfriendly

3.2.3 Scores for Affirmative and Negative Contexts

The sentiment value of a term may change in different contexts. For instance, negation is a frequent context that can modify the sentiment polarity or intensity of a particular word. While many studies process negation by reverting sentiment polarity (from positive to negative and vice-versa), others argue that sentiment reversion may not be adequate (Kiritchenko et al., 2014). For example, “frightening” is very negative but “not frightening” often suggests a less intense negative emotion.

To address negation more efficiently, we calculated sentiment scores for negated and affirmative (non-negated) contexts separately (Kiritchenko et al., 2014). We divided the training data set into an affirmative and a negated corpus. The negated set contains all negated contexts segments and the affirmative set is composed by the remaining segments. The negated contexts are the sentence segments starting with a negation word present in the Christopher Potts’ sentiment tutorial (<http://sentiment.christopherpotts.net/lingstruc.html>) and ending with one of the punctuation marks: ‘,’, ‘.’, ‘:’, ‘;’, ‘!’, ‘?’. Then, we create the stock market lexicon by applying the same procedure utilized for the “general” score (i.e., described in the previous subsection) on each corpus (affirmative and negated), producing two sentiment scores for each item that should be used in the respective context. However, some items do not have sufficient occurrences in each corpus in order to have both sentiment scores calculated. In such situations, we assign its “general” sentiment to the unavailable sentiment context score.

3.3 LEXICON EVALUATION

As explained in Section 3.2, we adopt two complementary evaluation procedures that use a time ordered training/test split: a single holdout (75%/25%) and a rolling window (with 20 windows, each with 2/3 for training and 1/3 for testing). The former

procedure creates a large lexicon that is publicly made available, while the latter procedure uses much less data to generate each lexicon but it allows us to get several test sets and thus execute statistical significance tests. For both evaluation methods, we performed SA in each test set by applying each lexicon.

For comparison purposes, we also executed SA in the same test sets using six large and popular lexicons:

- GI (Stone et al., 1966) – comprises around 11,000 words (http://www.wjh.harvard.edu/~inquirer/spreadsheet_guide.htm). In particular, we used all words of the “Positiv” and “Negativ” attributes.
- OL (Hu and Liu, 2004) – contains nearly 6,000 positive and negative terms. The lexicon also contains common social media misspelled words (<http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar>).
- MSOL (Mohammad et al., 2009) – classifies more than 75,000 n-grams as positive or negative (<http://saifmohammad.com/Lexicons/MSOL-June15-09.txt.zip>).
- MPQA (Wilson et al., 2005) – with around 8,000 entries (http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/). It contains a list of strong and weak subjective terms. We assigned half of the sentiment score (i.e., 0.5 or -0.5) to weaker terms.
- SWN (Baccianella et al., 2010) – with continuous sentiment values to the nearly 117,000 synsets (i.e., group of words that are semantically equivalent in some context) of the WordNet lexical database (<http://sentiwordnet.isti.cnr.it/downloadFile.php>). A word may have multiple scores, since it can belong to diverse synsets. To solve this issue, we averaged the positive and negative values for all (word, POS tag) pairs.
- FIN (Loughran and McDonald, 2011) – contains word lists commonly applied in financial text documents (http://www3.nd.edu/~mcdonald/Word_Lists.html). We adopted the terms classified as negative (2349) and positive (354).

The message overall sentiment value is computed as the sum of all its lexical scores. When lexicon bigrams are present in the text, we only sum the score of the bigrams and do not account for the score of their individual constituents. The message is classified as “bullish”, “bearish” or “neutral” according to the sign of the sum (positive, negative or zero). In SA applying lexicons with affirmative and negated scores, we also identified the affirmative and negated context segments in order to utilize the adequate sentiment score.

The classification measures used were:

- the percentage of correct classifications (CC_1);
- the percentage of unclassified messages, i.e., texts with no lexicon items (Unc);
- the percentage of correct classifications excluding unclassified messages (CC_2);
- precision for “bullish” (P_{Bull}) and “bearish” (P_{Bear}), given by $\frac{TP}{TP+FP}$, where TP denotes the number of true positives and FP the number of false positives;
- recall for “bullish” (R_{Bull}) and “bearish” (R_{Bear}), given by $\frac{TP}{TP+FN}$, where FN denotes the number of false negatives;
- F-score for “bullish” ($F_{1_{Bull}}$) and “bearish” ($F_{1_{Bear}}$), where $F_1 = 2 \frac{Precision * Recall}{Precision + Recall}$;
- macro-averaged F-score (F_{Avg}) that averages both F-scores ($F_{1_{Bull}}, F_{1_{Bear}}$).

We assume that the classification is correct when it matches the same sentiment (bullish or bearish) as provided by user who made the post. Under the rolling window scheme, we verified the statistical significance of CC_1 and F_{Avg} improvements obtained by a specific approach relatively to another one. The parametric paired Student’s t-test and the non-parametric Wilcoxon signed rank test were applied to pairs of lexicon creation methods.

3.4 CREATION OF INVESTOR SENTIMENT INDICATORS

In this work, we measured the correlation of microblogging sentiment indicators with two widely applied survey sentiment indicators: AAI (e.g., Fisher and Statman, 2000; Brown and Cliff, 2004) and II (e.g., Fisher and Statman, 2000; Brown and Cliff, 2004; Verma and Soydemir, 2009). AAI measures the percentage of individual investors who are bullish, bearish, and neutral based on the votes of their members to a poll questioning their sentiment on the stock market for the next six months. AAI values are published online each Thursday morning containing data from previous Thursday until last Wednesday. II analyzes each week over a hundred market newsletters and categorize each authors current opinion about the market as bullish, bearish or correction. The percentage of newsletters classified as bullish, bearish or correction are published every Wednesday and they include the newsletters analyzed until Tuesday. II measures may be more correlated to institutional sentiment than AAI, because many authors are market professionals (Brown and Cliff, 2004). AAI and II indicators were collected from Thompson Reuters Datastream (<http://online.thomsonreuters.com/datastream/>).

In the creation of microblogging sentiment indicators, we used Twitter for reasons of data availability. Twitter (twitter.com) is the most popular microblogging service. Unlike StockTwits, it is a generic platform. Yet, Twitter users also apply cashtags (i.e., "\$" followed by the respective ticker (e.g., \$GOOG)) in stock market conversations about those stocks. We collected all tweets containing cashtags of all stocks traded in US stock markets using Twitter REST API (<https://dev.twitter.com/docs/api>) and the R language (or statistical environment)². Twitter sentiment indicators were produced by applying SA on the collected Twitter data using two distinct lexicons:

² The total number of collected Twitter messages mentioning 3762 different cashtags is approximately 19 millions.

- TWTSML uses the selected Stock Market Lexicon (SML), i.e., the $\text{PMI}_{\text{BiScr}}$ lexicon created using 75% of StockTwits data (Section 3.5). Affirmative or negated context scores are applied in the respective segments.
- TWTSWN applies the SWN lexicon (the selected baseline lexicon, Section 3.5).

Six different values are computed for each time period:

$$\text{TWTSML}_{\text{bull},t} = \text{SML}_{\text{bull},t} / (\text{SML}_{\text{bull},t} + \text{SML}_{\text{bear},t}) \quad (33)$$

$$\text{TWTSML}_{\text{bear},t} = \text{SML}_{\text{bear},t} / (\text{SML}_{\text{bull},t} + \text{SML}_{\text{bear},t}) \quad (34)$$

$$\text{TWTSML}_{\text{spread},t} = \text{TWTSML}_{\text{bull},t} - \text{TWTSML}_{\text{bear},t} \quad (35)$$

$$\text{TWTSWN}_{\text{bull},t} = \text{SWN}_{\text{bull},t} / (\text{SWN}_{\text{bull},t} + \text{SWN}_{\text{bear},t}) \quad (36)$$

$$\text{TWTSWN}_{\text{bear},t} = \text{SWN}_{\text{bear},t} / (\text{SWN}_{\text{bull},t} + \text{SWN}_{\text{bear},t}) \quad (37)$$

$$\text{TWTSWN}_{\text{spread},t} = \text{TWTSWN}_{\text{bull},t} - \text{TWTSWN}_{\text{bear},t} \quad (38)$$

where:

- $\text{SML}_{\text{bull},t}$ corresponds to the sum of all positive SA scores (i.e., greater than zero) using SML (i.e., $\text{PMI}_{\text{BiScr}}$ lexicon) on all tweets from a given t time period.
- $\text{SML}_{\text{bear},t}$ corresponds to absolute value of the sum of all negative SA scores (i.e., less than zero) using SML on all tweets for time t .
- $\text{SWN}_{\text{bull},t}$ corresponds to the sum of all positive SA scores using the selected baseline lexicon (i.e., SWN lexicon) on all tweets for time t .
- $\text{SWN}_{\text{bear},t}$ corresponds to absolute value of the sum of all negative SA scores using the SWN lexicon on all tweets for time t .

We used survey and Twitter indicators from February 1, 2013 to March 27, 2015. This time period is subsequent to the applied in the creation of the selected stock market lexicon. Since both AAI and II use ratios, we also created bullish and bearish ratios. We decided to use SA scores instead of the number of bullish or bearish mes-

sages because we consider that they better indicate the sentiment strength. Additionally, we calculated the bull-bear spread ($TWTSML_{spread} - TWTSWN_{spread}$), a common measure of sentiment (e.g., Brown and Cliff, 2004; Verma and Soydemir, 2009). Different Twitter sentiment indicators (i.e., $TWTSML_{bull}$, $TWTSML_{bear}$, $TWTSML_{spread}$, $TWTSWN_{bull}$, $TWTSWN_{bear}$, $TWTSWN_{spread}$) were computed for each survey sentiment indicator corresponding to their time periods. Each Twitter sentiment indicator is correlated to their AAI and II counterparts (e.g., $TWTSML_{bull}$ with AAI_{bull} , $TWTSWN_{spread}$ with II_{spread}).

3.5 EXPERIMENTAL RESULTS AND DISCUSSION

3.5.1 Lexicon Evaluation

In this section we present the SA results for the tested lexicons. We start by analyzing the use of the proposed statistical measures in the computation of a unique sentiment score for each item. We experimented four different scores for the three main statistical measures (PMI, TFIDE, IG):

- *Scr* corresponds to the value calculated by the main statistical measure (S_{PMI} , S_{TFIDE} , S_{IG});
- *Assoc* is the product of *Scr* and M_{assoc} ;
- *Days* is the product of *Scr* and P_{days} ;
- *All* is the product of *Scr*, M_{assoc} and P_{days} .

Table 4 shows all classification metrics for the created lexicons using a unique context score and Table 5 indicates which lexicons obtain statistically significant higher CC_1 and F_{Avg} values compared with other lexicons according to the paired Student's t-test and the Wilcoxon signed rank test. The alternative hypothesis of these tests is that the lexicon in the row has higher results than the lexicon in the column.

Table 4.: Classification results for the created lexicons with unique context score (in %, best values in **bold**)

Lexicon	CC1	Unc	CC2	P _{Bull}	R _{Bull}	F1 _{Bull}	P _{Bear}	R _{Bear}	F1 _{Bear}	F _{Avg}
Panel A: Evaluation results of holdout split method										
PMI _{Scr}	75.2	0.5	75.5	88.6	76.5	82.1	51.9	71.5	60.1	71.1
PMI _{Assoc}	75.6	0.5	76.0	88.5	77.4	82.6	52.6	70.6	60.3	71.4
PMI _{Days}	78.8	0.5	79.1	86.0	85.4	85.7	59.5	59.6	59.6	72.6
PMI _{All}	78.8	0.5	79.2	86.0	85.5	85.8	59.7	59.5	59.6	72.7
TFIDF _{Scr}	74.3	0.5	74.7	88.6	75.1	81.3	50.6	71.9	59.4	70.4
TFIDF _{Assoc}	74.8	0.5	75.1	88.4	76.2	81.8	51.3	70.8	59.5	70.7
TFIDF _{Days}	78.4	0.5	78.7	85.6	85.4	85.5	58.8	58.3	58.6	72
TFIDF _{All}	78.5	0.5	78.8	85.5	85.5	85.5	59.1	58.1	58.6	72.1
IG _{Scr}	70.5	0.5	70.8	89.4	68.5	77.5	46.1	76.3	57.4	67.5
IG _{Assoc}	71.6	0.5	71.9	89.5	70.1	78.6	47.3	75.9	58.3	68.4
IG _{Days}	76.0	0.5	76.4	87.2	79.5	83.2	53.4	65.9	59.0	71.1
IG _{All}	76.4	0.5	76.7	86.9	80.4	83.5	54.1	64.9	59.0	71.3
Panel B: Average evaluation results of rolling window method										
PMI _{Scr}	71.1	0.5	71.4	90.0	69.0	78.0	45.7	77.0	56.9	67.4
PMI _{Assoc}	71.5	0.5	71.9	89.9	69.8	78.4	46.2	76.7	57.3	67.9
PMI _{Days}	77.3	0.5	77.7	87.4	81.5	84.3	54.1	64.4	58.5	71.4
PMI _{All}	77.5	0.5	77.8	87.3	81.7	84.4	54.6	64.2	58.7	71.5
TFIDF _{Scr}	71.6	0.5	71.9	89.1	70.8	78.8	45.6	73.4	55.8	67.3
TFIDF _{Assoc}	72.1	0.5	72.5	89.0	71.7	79.3	46.3	73.1	56.3	67.8
TFIDF _{Days}	77.1	0.5	77.5	86.4	82.4	84.3	54.0	60.5	56.6	70.5
TFIDF _{All}	77.3	0.5	77.7	86.5	82.7	84.5	54.5	60.4	56.9	70.7
IG _{Scr}	67.4	0.5	67.7	90.5	63.8	74.2	42.7	78.7	54.3	64.2
IG _{Assoc}	68.0	0.5	68.3	90.5	64.7	74.9	43.2	78.5	54.6	64.8
IG _{Days}	75.2	0.5	75.5	88.4	77.4	82.3	50.6	68.6	57.3	69.8
IG _{All}	75.5	0.5	75.8	88.4	77.8	82.6	51.0	68.5	57.6	70.1

Table 5.: Paired Student’s t-test and the Wilcoxon signed rank test for pairwise comparison of lexicons using unique context scores

	PMI _{Scr}	PMI _{Assoc}	PMI _{Days}	PMI _{All}	TFIDF _{Scr}	TFIDF _{Assoc}	TFIDF _{Days}	TFIDF _{All}	IG _{Scr}	IG _{Assoc}	IG _{Days}	IG _{All}
PMI _{Scr}	—								abcd	abcd		
PMI _{Assoc}	abcd	—							abcd	abcd		
PMI _{Days}	abcd	abcd	—		abcd	abcd	ab	ab	abcd	abcd	abcd	abcd
PMI _{All}	abcd	abcd	abcd	—	abcd	abcd	ab	ab	abcd	abcd	abcd	abcd
TFIDF _{Scr}					—				abcd	abcd		
TFIDF _{Assoc}	cd				abcd	—			abcd	abcd		
TFIDF _{Days}	abcd	abcd			abcd	abcd	—		abcd	abcd	acd	cd
TFIDF _{All}	abcd	abcd			abcd	abcd	abcd	—	abcd	abcd	abcd	acd
IG _{Scr}									—			
IG _{Assoc}									abcd	—		
IG _{Days}	abcd	abcd			abcd	abcd			abcd	abcd	—	
IG _{All}	abcd	abcd			abcd	abcd			abcd	abcd	abcd	—

The following symbols denote significance at the 5% level: a - paired Student’s t-test for F_{Avg} ; b - Wilcoxon signed rank test for F_{Avg} ; c - paired Student’s t-test for CC_1 ; d - Wilcoxon signed rank test for CC_1 . Alternative hypothesis: lexicon in the row has higher values than the lexicon in the column.

The best overall results (e.g., highest CC_1 , CC_2 and F_{Avg} values) are obtained by the PMI_{All} method, which delivers statistically significant higher CC_1 and F_{Avg} values than all other lexicons. The complementary metrics proved to be useful because they improved the evaluation results for all main statistical measures. Every lexicon applying M_{assoc} or P_{days} metrics obtain statistically significant higher CC_1 and F_{Avg} values than their Scr counterparts (e.g., PMI_{Assoc} , $TFIDF_{All}$ and IG_{Days} are higher than PMI_{Scr} , $TFIDF_{Scr}$ and IG_{Scr} , respectively). The M_{assoc} measure permits slight gains in both $F1_{Bull}$ and $F1_{Bear}$ scores. The P_{days} metric is able to substantially improve $F1_{Bull}$ values while maintaining the $F1_{Bear}$ very similar.

Next, we compare the selected PMI_{All} with diverse reference lexicons. The SA results obtained by these lexical resources are presented in Tables 6 and 7. PMI_{All} based lexicons achieve the best results for all evaluation metrics by a significant margin. For example, sentiment classification using this lexicon obtains a 18.2 (F_{Avg}), 18.7 (CC_2) and

21.4 (CC1) percentage point difference in the holdout split scheme when compared with the baseline resource that has the highest overall results (i.e., SWN). All these improvements are statistically significant. In addition, the approximately 20,000 lexical entries that belong to the lexicons created in this work are included in more messages (only 0.5% of the posts are not classified). In contrast, the generic SWN lexicon, which contains a larger number of lexical items (117,000), presents a higher unclassification rate (5.1%). FIN achieves the lowest F_{Avg} , CC1 and CC2 values, despite having the second highest P_{Bull} . The poorer FIN unclassified message performance (66%) confirms that there is a considerable difference between the lexical terms extracted from financial text documents and StockTwits messages. In effect, there are several popular StockTwits terms, such as “bearish”, “bullish”, “breakout”, “put” and “short”, that are not present in FIN. Since “bullish” and “bearish” are distinctive terms of stock market terminology, we verified their presence and classification in baseline lexicons. FIN and GI lexicons do not contain these terms and MSOL lexicon incorrectly classifies “bullish” as negative. The remaining lexicons assign the correct classification to these words.

Next, we analyze the differences between the selected large lexicon (PMI_{All} , 20550 items) and baseline lexicons (SWN, 117,000 entries). The lexicons are quite distinct, since only 2695 lexical terms (13% of PMI_{All}) belong to both lexicons. Indeed, the presence of diverse stock market terms in generic opinion lexicons is unlikely (O’Leary, 2011). Also, 42% of these common terms (1121) have different sentiment polarities, as shown in Table 8. In particular, Table 8 presents examples of terms associated with: stock price changes (e.g., dip, downside, explosive, outperform, rip, sink); stock expectations (e.g., overvalue, underestimate); and stock operations (e.g., long). Under non financial contexts, these terms can suggest different sentiment values. For example, “underestimate” is in general a negative verb but when related with stocks it can suggest an opportunity to buy. These differences highlight the importance of producing specialized stock market lexicons. For demonstration purposes, Figure 16 plots a word cloud of the most interesting PMI_{All} bullish and bearish terms. Diverse terms with different sentiment polarity or absent from SWN stand out in this figure.

Table 6.: Classification results for the selected lexicon creation method and baseline lexicons (in %, best values in **bold**)

Lexicon	CC1	Unc	CC2	P _{Bull}	R _{Bull}	F1 _{Bull}	P _{Bear}	R _{Bear}	F1 _{Bear}	F _{Avg}
Panel A: Evaluation results of holdout split method										
PMI _{All}	78.8	0.5	79.2	86.0	85.5	85.8	59.7	59.5	59.6	72.7
FIN	16.8	66.0	49.3	83.5	13.9	23.8	34.3	25.1	29.0	26.4
GI	37.7	25.8	50.8	82.5	36.2	50.3	37.5	42.1	39.7	45.0
MSOL	53.4	1.8	54.3	79.1	58.6	67.3	33.9	38.2	35.9	51.6
MPQA	36.9	37.5	59.0	80.6	40.6	54.0	34.3	26.3	29.8	41.9
OL	31.8	43.0	55.9	82.6	32.7	46.8	37.7	29.4	33.1	39.9
SWN	57.4	5.1	60.5	79.9	59.7	68.3	34.1	50.7	40.7	54.5
Panel B: Average evaluation results of rolling window method										
PMI _{All}	77.5	0.5	77.8	87.3	81.7	84.4	54.6	64.2	58.7	71.5
FIN	17.3	63.7	47.8	84.2	13.9	23.8	32.8	27.4	29.3	26.5
GI	37.4	25.8	50.4	82.6	36.1	50.2	35.8	41.1	37.8	44.0
MSOL	52.3	1.2	53.0	80.2	55.8	65.6	32.5	41.9	36.1	50.8
MPQA	39.1	34.7	59.8	81.3	42.6	55.8	35.2	28.4	31.1	43.5
OL	34.1	39.5	56.3	83.5	34.6	48.9	38.2	32.3	34.7	41.8
SWN	58.9	4.4	61.6	80.6	61.4	69.7	33.8	51.1	40.4	55.0

Table 7.: Paired Student’s t-test and the Wilcoxon signed rank test for pairwise comparison of PMI_{All} method with baseline lexicons

	FIN	GI	MSOL	MPQA	OL	SWN
PMI _{All}	abcd	abcd	abcd	abcd	abcd	abcd

The following symbols denote significance at the 5% level: a - paired Student’s t-test for F_{Avg}; b - Wilcoxon signed rank test for F_{Avg}; c - paired Student’s t-test for CCL; d - Wilcoxon signed rank test for CC1. Alternative hypothesis: lexicon in the row has higher values than the lexicon in the column.

Table 8.: Examples of lexical terms with different sentiment value in PMI_{All} and SWN

Item	POS tag	PMI _{All}	SWN	Item	POS tag	PMI _{All}	SWN
careful	adjective	negative	positive	overvalue	verb	negative	positive
dip	noun	positive	negative	outperform	verb	positive	negative
rip	verb	positive	negative	downside	noun	negative	positive
sink	verb	negative	positive	explosive	adjective	positive	negative
long	adjective	positive	negative	underestimate	verb	positive	negative

The utility of affirmative and negated context scores is assessed by comparing PMI_{All}, TFIDF_{All} and IG_{All} with the equivalent lexicons containing affirmative and negated context scores (PMI_{BiScr}, TFIDF_{BiScr}, IG_{BiScr}). Thus, all sentiment scores apply the complementary metrics (M_{assoc} and P_{days}). Tables 9 and 10 show the evaluation results.



Figure 16.: Bullish and bearish word cloud for a stock market lexicon (PMI_{All}).

The application of affirmative and negated context scores appears to be beneficial. Despite the reduced difference, lexicons having two context scores improve or maintain almost all evaluation results compared to the unique score counterparts. IG_{BiScr} obtains statistically significant higher $CC1$ and F_{Avg} values than IG_{All} and PMI_{BiScr} produces statistically significant higher $CC1$ values than PMI_{All} . Moreover, PMI_{BiScr} lexicon has statistically significant higher $CC1$ and F_{Avg} values than all IG and $TFIDF$ lexicons. Figure 17 shows the sentiment scores of all PMI_{BiScr} items for both contexts. We can observe that the sentiment reversion in negation is not always appropriate. Indeed, only 41% have their sentiment orientation modified in negated contexts. Moreover, many items have much stronger sentiment value in negated contexts than in affirmative contexts and vice-versa. For example, the term bearish has a -6.634 score in affir-

Table 9.: Classification results for unique and dual context scores (in %, best values in **bold**)

Lexicon	CC1	Unc	CC2	P _{Bull}	R _{Bull}	F1 _{Bull}	P _{Bear}	R _{Bear}	F1 _{Bear}	F _{Avg}
Panel A: Evaluation results of holdout split method										
PMI _{All}	78.8	0.5	79.2	86.0	85.5	85.8	59.7	59.5	59.6	72.7
TFIDF _{All}	78.5	0.5	78.8	85.5	85.5	85.5	59.1	58.1	58.6	72.1
IG _{All}	76.4	0.5	76.7	86.9	80.4	83.5	54.1	64.9	59.0	71.3
PMI _{BiScr}	79.0	0.5	79.3	86.2	85.4	85.8	59.8	60.3	60.1	73.0
TFIDF _{BiScr}	78.5	0.5	78.9	86.0	85.1	85.5	59.0	59.6	59.3	72.4
IG _{BiScr}	76.7	0.5	77.0	87.0	80.8	83.8	54.7	64.8	59.3	71.5
Panel B: Average evaluation results of rolling window method										
PMI _{All}	77.5	0.5	77.8	87.3	81.7	84.4	54.6	64.2	58.7	71.5
TFIDF _{All}	77.3	0.5	77.7	86.5	82.7	84.5	54.5	60.4	56.9	70.7
IG _{All}	75.5	0.5	75.8	88.4	77.8	82.6	51.0	68.5	57.6	70.1
PMI _{BiScr}	78.3	0.5	78.7	86.4	84.2	85.2	56.8	60.0	58.1	71.7
TFIDF _{BiScr}	77.3	0.5	77.7	86.5	82.6	84.4	54.4	60.7	57.0	70.7
IG _{BiScr}	75.6	0.5	76.0	88.6	77.8	82.7	51.2	69.2	58.0	70.3

mative contexts and just -0.794 in negated contexts. For instance, saying “not bearish” does not signify the same as being bullish. Usually it just means that the opinion is not pessimistic. In the opposite situation, bailout has -5.392 points for negated contexts and only -0.657 for affirmative segments. The refusal of a bailout may imply the business downfall while its application may not necessarily mean a successful future. Negation handling is not a straightforward procedure, it may vary according to each term. Thus, the use of two context scores may be very useful in this matter. The PMI_{BiScr} lexicon created using the first 75% labeled messages is available at https://github.com/nunomroliveira/stock_market_lexicon.

3.5.2 Correlation with Survey Sentiment Indicators

To evaluate the relevance of microblogging sentiment indicators created using the stock market lexicon, we assess the association between Twitter sentiment indicators and two popular survey sentiment indicators: AAI and II. A strong correlation may indicate that the microblogging sentiment indicator can be an acceptable alternative or proxy.

Table 10.: Paired Student’s t-test and the Wilcoxon signed rank test for pairwise comparison of lexicons using unique and two context scores

	PMI _{All}	TFIDF _{All}	IG _{All}	PMI _{BiScr}	TFIDF _{BiScr}	IG _{BiScr}
PMI _{All}	—	ab	abcd		ab	abcd
TFIDF _{All}		—	acd			cd
IG _{All}			—			
PMI _{BiScr}	cd	abcd	abcd	—	abcd	abcd
TFIDF _{BiScr}			acd		—	cd
IG _{BiScr}			abcd			—

The following symbols denote significance at the 5% level: a - paired Student’s t t-test for F_{Avg} ; b - Wilcoxon signed rank test for F_{Avg} ; c - paired Student’s t t-test for $CC1$; d - Wilcoxon signed rank test for $CC1$. Alternative hypothesis: lexicon in the row has higher values than the lexicon in the column.

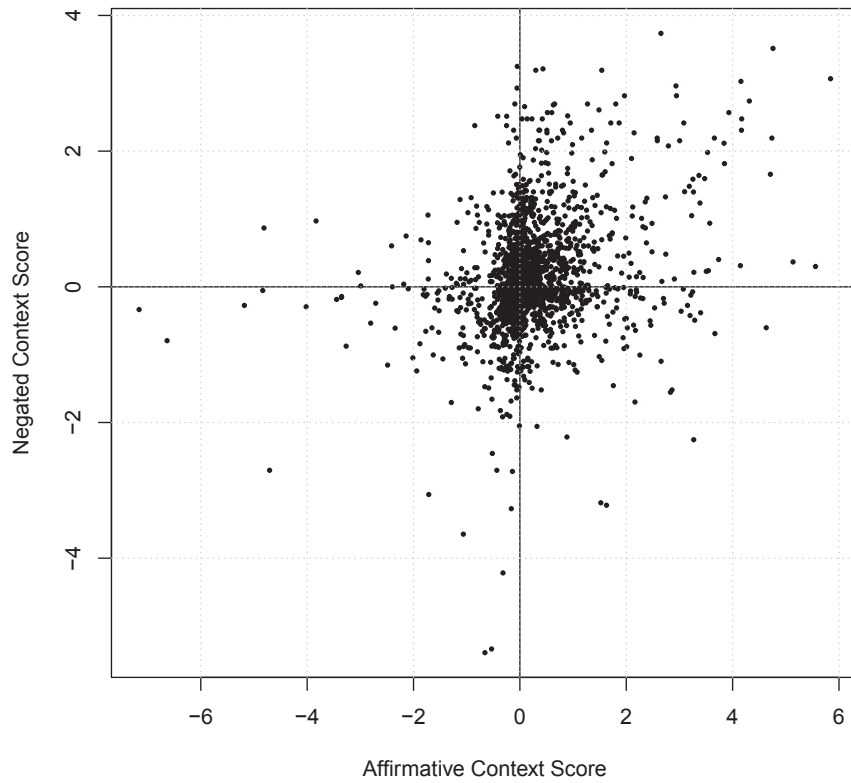


Figure 17.: Distribution of sentiment scores of SML for affirmative and negated contexts

The correlation calculation uses 112 observations for AAll and 110 observations for All. Table 11 presents the respective Pearson’s correlation values.

Table 11.: Pearson’s correlation values of Twitter sentiment indicators with survey sentiment indicators (\star – p-value < 0.01 , \diamond – p-value < 0.05 , best correlation values for each survey sentiment value in **bold**)

Pair	Correlation	Pair	Correlation
(TWTSML _{bear} , AAI _{bear})	0.489 \star	(TWTSWN _{bear} , AAI _{bear})	0.436 \star
(TWTSML _{bull} , AAI _{bull})	0.233 \diamond	(TWTSWN _{bull} , AAI _{bull})	0.220 \diamond
(TWTSML _{spread} , AAI _{spread})	0.376 \star	(TWTSWN _{spread} , AAI _{spread})	0.342 \star
(TWTSML _{bear} , II _{bear})	0.540 \star	(TWTSWN _{bear} , II _{bear})	0.628 \star
(TWTSML _{bull} , II _{bull})	0.533 \star	(TWTSWN _{bull} , II _{bull})	0.445 \star
(TWTSML _{spread} , II _{spread})	0.585 \star	(TWTSWN _{spread} , II _{spread})	0.551 \star

The obtained results show that Twitter sentiment indicators have a statistical significant moderate correlation with diverse survey sentiment values. Indeed, only the bullish value of AAI is poorly correlated with both Twitter sentiment indicators. II indicators present higher correlation values than AAI, so it may indicate that Twitter users posting about stock market are informed traders because II is more associated to professional investors than AAI (Brown and Cliff, 2004). Moreover, sentiment indicators produced with the selected stock market lexicon (SML, i.e., PMI_{BiScr}) are more correlated to almost all survey sentiment values than indicators created with the selected baseline lexicon (i.e., SWN). Only II_{bear} is less correlated with TWTSML values than with TWTSWN values.

Sentiment indicators created using automated computational methods have various advantages regarding the traditional sentiment indicators produced from surveys. For instance, the creation of these sentiment indicators is faster and cheaper, permits higher frequencies (e.g., daily) and may be targeted to a more restricted set of stocks (e.g., stock market indices or individual stocks). Therefore, the application of SA in microblogging data may constitute a valuable alternative to the creation of investor sentiment indicators. Additionally, the utilization of stock market lexicons allows an easy and fast unsupervised production of these indicators.

3.6 CONCLUSIONS

With the expansion of social media (e.g., Twitter, message boards), the interest in SA as increased, allowing the summary of opinions from large amounts of opinionated messages and thus support decision-making in several domains, including stock markets. A sentiment lexicon is a crucial resource for SA, enabling an easy, fast and accessible unsupervised SA and avoiding the expensive and arduous task of manually labeling data. Moreover, opinion lexicons permit the creation of very informative features for supervised SA. However, there are very few financial lexicons (e.g., Loughran and McDonald, 2011; Mao et al., 2014) and the existing domain independent lexicons (e.g., Stone et al., 1966; Wilson et al., 2005; Baccianella et al., 2010) may not be adjusted to the stock market domain.

In this study, we propose an automated and fast approach to create stock market lexicons for microblogging messages. We employed a large labeled data set of Stock-Twits messages and tested three adaptations and two novel statistical measures to calculate the sentiment score. Also, we suggest the use of sentiment scores for affirmative and negated contexts in order to improve the difficult task of negation processing. The results on the test data confirmed that these newly created lexicons substantially increase the SA when compared with six reference lexicons. The improvements in evaluation metrics obtained by the created lexicons are statistically significant. Furthermore, the use of the proposed complementary metrics proved to be useful. Lexicons applying any of these measures obtain statistically higher evaluation results in SA than their counterparts that do not use the complementary metrics. Moreover, the utilization of affirmative and negated context scores appears to be beneficial. Lexicons applying these measures improve or maintain almost all evaluation results compared to their counterparts. Some of these improvements are statistically significant. A substantial contribution of this work is to make publicly avail-

able a large stock market lexicon with context scores. This is accessible at: https://github.com/nunomroliveira/stock_market_lexicon.

Also in this work, we selected a stock market lexicon (SML, i.e., PMI_{BIScr}) and a baseline lexicon (SWN) to easily generate investor sentiment indicators from Twitter messages holding cashtags of stocks traded in US markets. Twitter based sentiment indicators showed a significant moderate Pearson's correlation with the widely applied AAI and II survey sentiment indicators. Therefore, the Twitter based sentiment indicator can be used as an acceptable proxy for survey sentiment indicators. Moreover, the sentiment indicators created with the proposed lexicon showed higher correlations values than indicators produced with the baseline lexicon in five of the six analyzed survey indicators. A microblogging sentiment indicator presents several advantages when compared with survey sentiment indicators: it is faster and cheaper to produce, it allows higher frequencies (e.g., daily) and it can be adjusted to both stock market indices and individual stocks.

The proposed procedure allows the fast and effortless creation of a lexicon properly adapted to stock market contents. This lexicon may permit an easy and effective unsupervised SA related to the stock market domain, such as the creation of investor sentiment indicators.

Our results suggest that the proposed microblogging sentiment lexicon approach might be a useful source of information for stock market participants, and this merits future research. For instance, a collective intelligence approach can be used to more easily assign sentiments to unlabeled text and identify stock market terms, thus widening the applicability of the proposed procedure to other stock market message sources (e.g., Twitter) and producing more accurate and comprehensive lexicons. Active learning algorithms may complement this approach by automatically selecting a reduced but more relevant set of text messages for human classification, thus reducing the manual labeling effort. Moreover, it is important to analyze the informative content of microblogging sentiment indicators to forecast stock market behavior. Sentiment indicators created by SA using stock market lexicons can be included in models

to predict diverse stock market variables (e.g., returns, trading volume, volatility) and survey indicators in order to assess their predictive ability. This study shows that the created lexicon permits an easy, fast and light production of accurate investor sentiment indicators from microblogging data. This research line is addressed in the next chapter.

THE IMPACT OF MICROBLOGGING DATA FOR STOCK MARKET PREDICTION

4.1 INTRODUCTION

Due to the growth of the Internet and Web 2.0 phenomenon, social media is an important big data source (Fan and Gordon, 2014). Users spend a significant part of their time on social media services. Thus, the analysis of these social media data may allow a deeper understanding of users' behavior that can be utilized for various purposes, including the financial domain. For instance, Thomson Reuters Eikon and Bloomberg are examples of financial services that include SA of tweets^{1,2}.

In effect, the usage of sentiment and attention indicators for stock market behavior modeling and prediction is an active research topic. As presented in Subsection 2.4.3 and summarized in Table 2, the literature of this topic can be distinguished in terms of several dimensions. For instance, several proxies for sentiment have been used in these studies (e.g., surveys, financial data, message boards, news, microblogs, blogs). In particular, textual contents (e.g., microblogs, message boards) permits a faster, cheaper and more flexible creation of these sentiment indicators compared to surveys. For ex-

¹ <http://financial.thomsonreuters.com/en/products/tools-applications/trading-investment-tools/eikon-trading-software.html> [Accessed on 5 December 2016]

² <http://www.bloomberg.com/company/announcements/trending-on-twitter-social-sentiment-analytics/> [Accessed on 5 December 2016]

ample, it allows the production of indicators having higher periodicities (e.g., intraday, daily) and targeted to a specific group of stocks.

The extraction of sentiment indicators from text is based on two main approaches: supervised and unsupervised. Yet, the utilization of supervised ML (e.g., NN, SVM) needs labeled training data that is often unavailable. Moreover, it can be computationally very expensive, requiring unaffordable computer resources and excessive training periods. Therefore, many studies opt for unsupervised methods based on lexicons or keywords (Bollen et al., 2011; Mao et al., 2011). The majority of applied lexicons are generic (e.g., GI, MPQA, SWN) and only two studies apply a financial lexicon created by Loughran and McDonald (2011). However, as explained in Chapter 3, domain independent lexicons are ineffective for stock market messages and the financial lexicon of Loughran and McDonald (2011) obtains low recall values for microblogging messages. For instance, the mentioned financial lexicon do not include popular stock markets terms such as “bearish”, “bullish”, “breakout”, “put” and “short”. Thus, in this study we use the extensive lexicon adjusted to microblogging stock market conversations created in Chapter 3 that should produce more accurate sentiment values. We opted for a lexicon-based SA approach because we obtained good results in the previous experiments described in Chapter 3. Moreover, the application of state of the art ML methods (e.g., deep learning, SVM) requires high computational effort to classify the several millions of microblogging messages applied in this project (e.g., 31 million tweets). The utilization of a simple rule based SA using the created stock market lexicon produces satisfactory results in a much faster and accessible procedure.

The diverse existing investor sentiment indicators are extracted from different sources (e.g., surveys, social media) and have distinct frequencies (e.g., monthly, weekly, daily). Therefore, they usually have different characteristics and may contain some noise. The application of a KF procedure to aggregate diverse sentiment indicators may permit the creation of a less noisy and more representative sentiment indicator than their individual components. KF was already used to combine sentiment based on financial data and surveys (Brown and Cliff, 2004). We believe that there were no previous stud-

ies aggregating social media with other sources, particularly for higher periodicities than the weekly. Thus, we experiment the creation of an unique daily indicator from microblogging data and various weekly and monthly survey sentiment indicators by using a KF procedure.

The dimension of the data sets applied in this topic is usually high for studies using financial data, surveys or news but it is low for social media data. In this project, we apply almost three years of Twitter data while studies using social media data use less than two years of data.

The analysis of the impact of sentiment on portfolios based on some characteristics (e.g., size, book to market, volatility) is common on financial data or surveys studies. Nevertheless, we did not find any evaluation of the influence of sentiment created from social media on portfolios of any type. Furthermore, this analysis is mainly performed for lower periodicities than the daily. Therefore, we study the predictive value of daily sentiment extracted from Twitter data on returns of portfolios formed on industries and size.

Most studies apply MR and VAR models to assess the informative content of sentiment for stock market behavior. The application of more flexible learning ML models (e.g., SVM (Deng et al., 2011; Nguyen et al., 2015; Schumaker et al., 2012), NN (Bollen et al., 2011)) is limited. Furthermore, only two studies compared the results of different ML models for stock market variables (Oh and Sheng, 2011; Deng et al., 2011). In this study, we compare five regression models, MR, NN, SVM, RF and EA, applying a more robust rolling window validation scheme that was only used in one related work (Deng et al., 2011).

The majority of the studies presents limitations in terms of lack of a robust evaluation. For example, most studies did not use a out of sample evaluation and other works applied very short test sets: 10 predictions (Oh and Sheng, 2011); 19 forecasts (Bollen et al., 2011); 20 and 30 forecasts (Mao et al., 2011); 25 and 35 predictions (Chen and Lazer, 2013); and 79 forecasts (Nguyen et al., 2015). Additionally, the application of statistical tests to evaluate the predictive accuracy is very scarce (Table 2). This study

applies the DMST test for predictive accuracy to evaluate the statistical significance of the predictions.

The forecasting of survey sentiment indices can be valuable for investors because it may allow an anticipation of their values or constitute an inexpensive alternative measure. Some studies perform regressions of survey sentiment values applying contemporaneous values of other sources (e.g., financial measures, other surveys) (Brown and Cliff, 2004; Schmeling, 2007). However, we did not find their forecasting using lagged values of other sentiment sources, specially social media sentiment. Therefore, we forecast two popular survey sentiment indicators (AAII and II) using Twitter and KF sentiment indicators.

Regarding the obtained results, diverse studies found text based sentiment useful to decide trading strategies (Schumaker et al., 2012; Al Nasser et al., 2015) or forecast stock market variables (e.g., stock prices (Bollen et al., 2011), price directions (Nguyen et al., 2015), returns (Tetlock, 2007; Sabherwal et al., 2011), volatility (Sabherwal et al., 2011), trading volume (Antweiler and Frank, 2004)). Sentiment extracted from text has been almost exclusively applied in the forecasting of daily variables of individual stocks, indices or aggregate market. Nonetheless, sentiment based on surveys and financial data has been mainly used for lower periodicities (e.g., monthly) and portfolios (e.g., firm dimension, age, volatility, book-to-market ratio). Sentiment seems to have more impact on returns of some portfolios having extreme values on these characteristics (Neal and Wheatley, 1998; Brown and Cliff, 2005; Baker et al., 2012). We believe there are no studies applying microblogging data to forecast variables based on portfolios, as executed in this work. Posting volume on social media services has also been used to forecast volatility (Antweiler and Frank, 2004; Das and Chen, 2007) or trading volume (Sprenger et al., 2014).

However, these findings are not consensual. The efficient market hypothesis proposes that investors act as rational agents and all available information is reflected immediately in stock prices, which implies that assets are traded at their fair value (Fama, 1998). Additionally, there are studies that find little evidence of the predictive value

of sentiment for stock prices or returns (e.g., Tumarkin and Whitelaw, 2001; Antweiler and Frank, 2004; Timmermann, 2008; Brown and Cliff, 2004) and posting volume for volatility (Antweiler and Frank, 2004; Tumarkin and Whitelaw, 2001).

The main goal of this study is to assess the value of microblogging data to the forecasting of stock market variables. To achieve this objective, we propose the methodology that is presented in Figure 18 and detailed in the next sections. Our predictive models are fitted using microblog sentiment and attention indicators. These indicators are based on Twitter (Twitter sentiment indicators (TWT)) and on weekly (AAII, II) and monthly (UMCS, Sentix) survey indices retrieved using the Datastream service. A fast unsupervised SA is performed by using a specialized financial microblog lexicon created using the procedure described in Chapter 3. This is the first study that adopts a specialized stock market lexicon adapted to microblogging messages. The individual tweet and survey sentiment indicators are aggregated into daily values using diverse forms (e.g., Bullish Ratio (BullR)). We also use KF to merge weekly and monthly survey indices with the daily microblog sentiment indicators, which is a new approach in this context, and test its forecasting ability. Using rolling windows, we fit five distinct ML models to predict daily stock indices and portfolio values (e.g., SP500, Return of the Market minus the Risk-Free return rate (RMRF)) that were collected from Datastream and Prof. Kenneth French Webpage. We also fit five ML models to predict weekly sentiment survey indices (AAII and II). To our knowledge, this is the first attempt to predict these survey indices. Finally, we assume that predictions using sentiment and attention indicators are valuable when they are statistically better, according to the DMST test, than an Auto-Regressive (AR) baseline model.

Our main contributions are summarized as follows:

- 1) we propose a robust methodology to assess the value of text based sentiment and attention indicators when forecasting stock market variables (Figure 18). The methodology assumes a realistic rolling window evaluation (with 300 training days for stock market variables and 50 observations for survey indicators), the

Chapter 4. The Impact of Microblogging Data for Stock Market Prediction

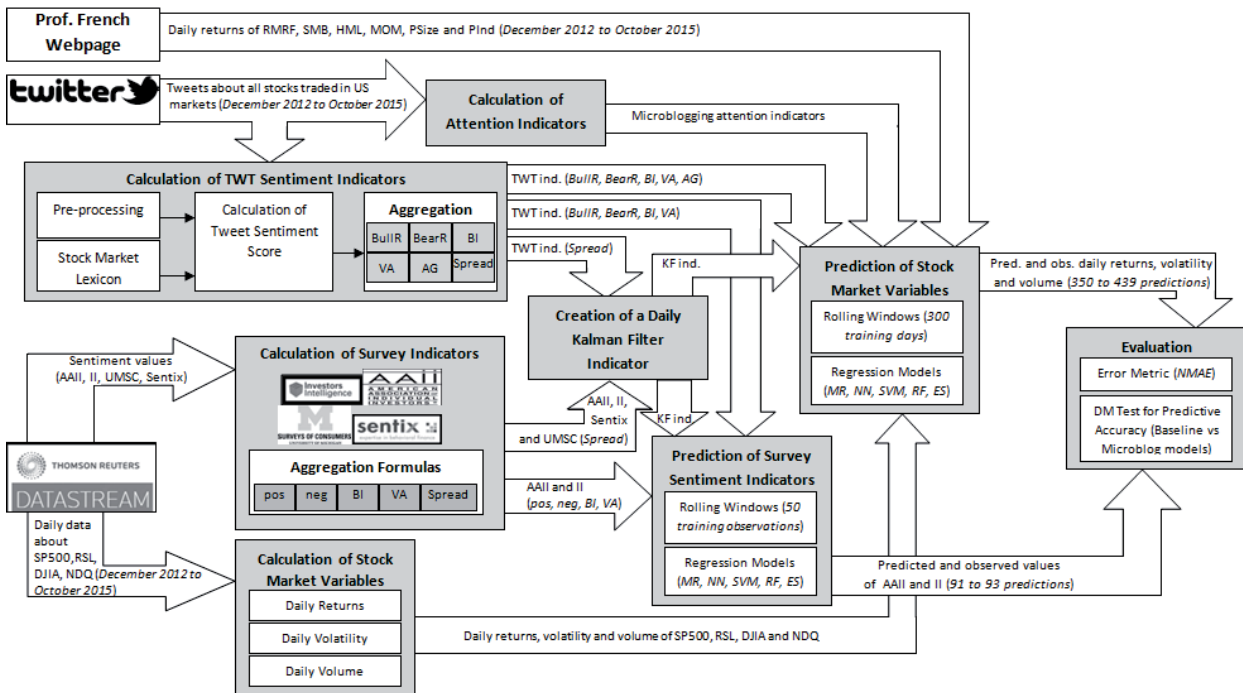


Figure 18.: Schematic of the proposed methodology

DMST test for predictive accuracy and five ML regression models (MR, NN, SVM, RF and EA), under two main strategies: baseline model (without microblog features) and microblog based (with such features). This methodology can be easily adapted in future research on this topic;

- i) we generate new Twitter sentiment indicators based on a recent lexicon (Oliveira et al. (2016) and Chapter 3) that contains more than 20,000 entries and that is specifically adjusted to financial microblogs. These indicators were extracted from a recent and very large Twitter based data set, containing around 31 million messages from December 2012 to October 2015 related with all stocks traded in US markets (about 3,800 stocks);
- ii) we propose a less noisy KF sentiment indicator that combines measures of different periodicities: daily Twitter indicators, weekly AAII and II values and monthly UMCS and Sentix indices. We also explore diverse sentiment aggregation formulas (e.g., BullR, Variation (VA));

- iv) we conduct a large set of experiments, predicting daily stock market variables (returns, trading volume and volatility) of diverse indices, such as SP500 and NDQ, and portfolios formed on size and industries;
- v) we also predict two popular survey sentiment indicators (AAII and II) using Twitter and KF sentiment indicators, which may permit a satisfactory anticipation of AAII and II values or a decent alternative whenever they are unavailable.

We consider that this study may support further advances in financial decision support systems. For instance, real time financial systems can be enhanced by providing personalized sentiment and prediction values related to specific stocks (e.g., individual stocks, indices) or surveys (e.g., AAII, II) for diverse periodicities (e.g., intraday, daily, weekly). Thus, it may contribute to the creation of more flexible financial systems that are more capable to satisfy user needs.

This chapter is structured as follows. Section 4.2 describes the creation of microblogging sentiment and attention indicators. Section 4.3 presents the applied survey sentiment indicators. Section 4.4 explains the KF procedure used to create an unique daily sentiment indicators combining daily microblogging indicators and weekly and monthly survey sentiment. Section 4.5 characterizes the stock market data applied in this study. Section 4.6 presents the predictive models for stock market variables and survey values. Section 4.7 describes the evaluation scheme utilized to assess the predictive value of microblogging data for stock market behavior and survey sentiment. Section 4.8 shows and analyzes the results of the prediction of returns, volatility, trading volume and survey sentiment values. Finally, Section 4.9 presents and discusses the main conclusions of this study.

4.2 MICROBLOGGING SENTIMENT AND ATTENTION INDICATORS

Microblogging data have characteristics that may indicate potential value to the forecasting of stock market behavior. Services such as Twitter and StockTwits have large

communities of investors sharing information about stock market. These users frequently interact during the day and react readily to events. Messages are usually very objective due to the character limit. Microblogging users generally apply cashtags in stock market conversations to refer to the involved stocks. Cashtags are composed by a "\$" character and the respective ticker (e.g., \$AAPL) and its presence means that the message is related with that stock. The utilization of cashtags permits an easy and less noisy selection of messages related to specific stocks. The extraction of attention and sentiment from microblogging is faster and cheaper than from traditional sources (e.g., surveys) because data is promptly available at very low cost.

The sentiment and attention indicators created in this study were extracted from Twitter, which is a large microblog platform with more than 300 million active users³. Using Twitter REST API (<https://dev.twitter.com/docs/api>), we collected all messages (around 31 million) from 22nd of December 2012 to 29th of October 2015 holding cashtags of all stocks traded in US markets (nearly 3,800 stocks). R tool (<http://www.r-project.org>) was used in all processing tasks and MongoDB (<https://www.mongodb.org>) was applied to store Twitter data.

The number of tweets was applied in the production of the attention indicators. We opted to use the first difference of the posting volume because there is a visible growing number of tweets during the analyzed time period. To create the investor sentiment indicators, we used the sentiment scores produced by SA on all tweets. SA applies a recently proposed lexicon (Oliveira et al. (2016) and Chapter 3) that is properly adapted to microblogging conversations about stock market and it is publicly available at https://github.com/nunomroliveira/stock_market_lexicon. This lexicon was automatically created using data from June, 2010 to March, 2013. It is an up-to-date lexicon because its training data ends nearly one year before the first prediction day of this study. It contains approximately 7000 unigrams, 13000 bigrams and the respective sentiment scores for affirmative and negated contexts. For instance, a negative score indicates a bearish word and a positive sentiment value indicates a bullish

³ <http://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/> [Accessed on 7 December 2016]

word. Negated contexts are text segments starting with a negation expression while the affirmative contexts are all other segments. To identify affirmative and negated segments, we applied the same approach used in the lexicon creation. Negated segments begin with a negation item included in the Christopher Potts' sentiment tutorial (<http://sentiment.christopherpotts.net/lingstruc.html>) and end with one of the following punctuation marks: ",", ".", ":", ";", "!", "?". The sentiment score of each tweet corresponds to the sum of the sentiment value of all lexicon items present in the message. When lexicon bigrams are identified in the text, we only account the score of the bigrams and do not consider the score of their individual components. In our opinion, bigrams scores are more precise than unigram scores because bigrams have a more defined context. For instance, the bigram "debt free" is usually bullish while its individual components may have distinct sentiment orientation (e.g., "greek debt", "more debt", "free fall"). To adequately verify the presence of lexicon elements, we executed the preprocessing tasks:

- replace all cashtags by the tag "tkr"; all numbers by the tag "NUM"; all mentions by "@user"; all URL addresses by "URL";
- execute tokenization, POS tagging and lemmatization by applying Stanford CoreNLP (Toutanova et al., 2003).
- identify the affirmative and negated segments in order to apply the adequate score.

The sentiment indicators are created using the scores produced by the SA. We created two major types of investor sentiment indicators: general and sectorial. The general indicators represent the sentiment of the whole investor community. Thus, we used all tweets in the construction of these indicators. The sectorial indicators measure the sentiment regarding specific sectors (e.g., industries). In the creation of these indicators, we applied tweets enclosing cashtags of stocks belonging to the respective sector.

We selected the cashtags composing each sector based on their Standard Industrial Classification (SIC) code.

We also experimented diverse forms to calculate the daily sentiment values:

- BullR (Oliveira et al., 2013):

$$BullR_t = \frac{NBull_t}{NBull_t + NBear_t} \quad (39)$$

- Bearish Ratio (BearR):

$$BearR_t = \frac{NBear_t}{NBull_t + NBear_t} \quad (40)$$

- Bullishness Index (BI) (Antweiler and Frank, 2004; Sprenger et al., 2014):

$$BI_t = \ln \frac{NBull_t + 1}{NBear_t + 1} \quad (41)$$

- VA (Oliveira et al., 2013):

$$VA_t = BullR_t - BullR_{t-1} \quad (42)$$

- AG (Antweiler and Frank, 2004; Sprenger et al., 2014):

$$AG_t = 1 - \sqrt{1 - \left(\frac{NBull_t - NBear_t}{NBull_t + NBear_t} \right)^2} \quad (43)$$

where $NBull_t$ and $NBear_t$ are the bullish and bearish score of day t . We did not find any paper applying a sentiment measure computed exactly like BearR, however there are papers using similar measures, such as (Tetlock, 2007). The tested sentiment aggregation measures are distinct formulas applied in diverse studies in this research topic. In this study, we applied BullR, BI and VA indicators in the prediction of returns,

volatility, trading volume and survey sentiment values. Since dispersion of expectations is considered to be related with trading volume (Shalen, 1993; Antweiler and Frank, 2004) and volatility (Shalen, 1993), we added AG indicators in the forecasting of these stock market variables. Additionally, we applied BearR indicators in the prediction of negative values of survey sentiment values, because we used Twitter indicators counterparts in the forecasting of each survey sentiment value (e.g., BullR indicators applied in the prediction of AAI positive values or BearR indicators in the forecasting of AAI negative values).

4.3 SURVEY SENTIMENT INDICATORS

Survey sentiment indicators are frequently applied in studies about the analysis of the sentiment impact on stock market behavior. For instance, AAI provides weekly values of the votes of their members to a poll questioning their sentiment (bullish, bearish, neutral) on the stock market for the next six months. Research works such as (Fisher and Statman, 2000; Verma and Soydemir, 2009) used AAI index as a sentiment measure. Also, UMCS is a monthly sentiment index constructed from a consumer confidence survey answered by a random group of five hundred continental US households. Despite being considered less related to investor sentiment, UMCS index has been applied in stock market studies (e.g., Fisher and Statman, 2000). Moreover, Sentix (www.sentix.de) creates sentiment indices for various stock markets (e.g., US, Japan, Germany, Euro zone) from surveys answered by more than 3,500 participants (Schmeling, 2007). Another example is the II weekly index based on over a hundred independent market newsletters, with each newsletter being categorized as bullish, bearish or correction. II measures may be more correlated to institutional sentiment than AAI, because many of the newsletters' authors are market professionals (Brown and Cliff, 2004). The II index is widely applied in behavior finance literature (e.g., Solt and Statman, 1988; Fisher and Statman, 2000; Verma and Soydemir, 2009).

This study tests the prediction of weekly sentiment indicators (i.e., AAI, II) by using Twitter sentiment measures. This procedure may allow an acceptable anticipation of AAI and II values or a satisfactory alternative whenever they are unavailable. We did not experiment the forecasting of monthly sentiment indicators because there is no sufficient data to perform a feasible analysis. However, we also used monthly sentiment measures (i.e., UMCS, Sentix for US) in the creation of a sentiment indicator using the KF measure described in the next subsection. AAI, II, UMCS and Sentix values were obtained from Thompson Reuters Datastream (<http://online.thomsonreuters.com/datastream/>).

4.4 SENTIMENT INDICATORS CREATED BY KALMAN FILTER PROCEDURE

KF permits the combination of diverse observed variables in order to extract a latent variable. Existing investor sentiment indicators are measured by different approaches at different frequencies (e.g., surveys or social media interactions) and they usually produce distinct values. These observed values are related to sentiment but they contain some noise. The utilization of KF may permit the creation of a sentiment indicator that is more representative and less noisy than their individual components. Moreover, KF allows the usage of indicators with distinct frequencies (e.g., daily, weekly, monthly). Thus, we can produce a daily sentiment indicator from the combination of daily, weekly and monthly values. The linear dynamic model can be represented as follows:

$$\begin{aligned} Y_t &= F_t \theta_t + v_t, & v_t &\sim N(0, V_t) \\ \theta_t &= G_t \theta_{t-1} + w_t, & w_t &\sim N(0, W_t) \end{aligned} \tag{44}$$

where the first equation describes the observed variables (e.g., survey or social media sentiment indicators), the second equation represents the latent variable and V, W are parameter matrices. θ_0 is assumed to be normally distributed with mean 0 and variance $1e7$. The model is estimated by maximum likelihood allowing the observation

noises, v_t , to be cross-correlated. To reduce the complexity of the optimization problem and ensure that the variance-covariance matrix is positive semi-definite we use the approach suggested by Pinheiro and Bates (1996) and followed by Petris (2010), parametrizing the covariance matrix, V , in terms of the elements of its log-Cholesky decomposition and the system variance, W , using its log.

We created a daily sentiment indicator by applying the KF procedure to five different sentiment indicators: AAIL, II, UMCS, Sentix and the daily TWT created in this work. AAIL, II and TWT values correspond to the Bull-Bear spread (e.g., Verma and Soydemir, 2009):

$$SP_t = \frac{NBull_t - NBear_t}{NBull_t + NBear_t} \quad (45)$$

All indicators were normalized by calculating their standard score. The model parameters were estimated using the first training rolling window also applied in the forecasting of stock market variables (i.e., first 300 days). Then, we created the sentiment indicators for the entire time period by filtering the series using the estimated model.

KF values were tested as sentiment indicators in forecasting models to assess eventual improvements compared to the utilization of TWT as proxy for sentiment. As an example, Figure 19 shows the overall (for all analyzed stocks) KF sentiment indicator values and its individual components (AAIL, II, UMCS, Sentix, TWT) when considering the period that ranges from 1st January of 2014 to 30th June of 2014. The plot confirms that KF indicator presents smoother values than TWT.

4.5 STOCK MARKET DATA

Various studies have analyzed the influence of sentiment on portfolios formed on different characteristics (e.g., market capitalization, book-to-market ratio, industries). For example, some papers refer that sentiment has more impact on returns of stocks with lower market capitalization (Baker and Wurgler, 2006; Baker et al., 2012). However,

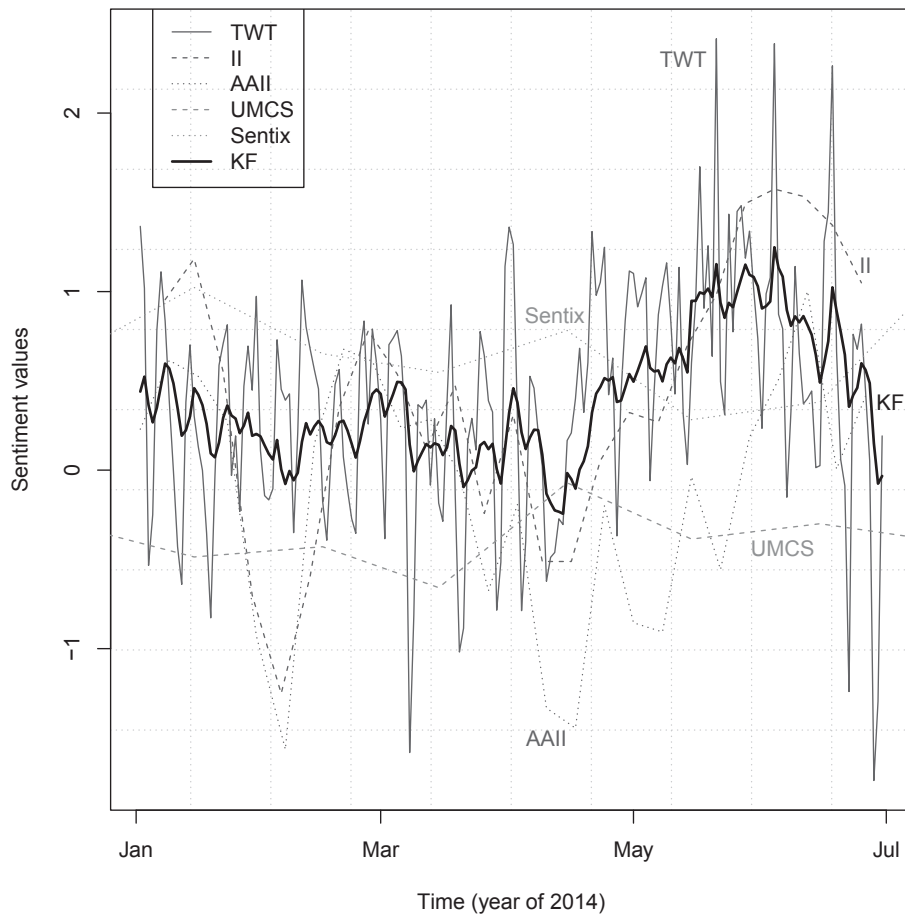


Figure 19.: Sentiment indicator values (KF, AAIL, II, UMCS, Sentix and TWT)

few of these studies apply sentiment measures extracted from social media and higher frequencies than the monthly periodicity. Therefore, in this study we explored a comprehensive set of stocks and portfolios having distinct characteristics:

- SP500: index composed by 500 large companies.
- Russell 2000 (RSL): index of the smallest 2000 companies belonging to Russell 3000 index.
- Dow Jones Industrial Average (DJIA): constituted by 30 large companies listed on NYSE and Nasdaq.
- NDQ: includes the 100 of the largest non-financial stocks traded on Nasdaq.

- RMRF: return of the market minus the risk-free return rate. The return of the market corresponds to the value-weight return of all Center for Research in Security Prices companies integrated in the US and traded on the NYSE, American Stock Exchange, or Nasdaq while the risk-free return rate is the one-month Treasury bill rate.
- Small Minus Big (SMB): a Fama and French factor corresponding to the difference in returns between small and large firms.
- High Minus Low (HML): a Fama and French factor that is equal to the difference in returns between value (i.e., high book-to-market ratios) and growth (i.e., low book-to-market ratios) stocks.
- Momentum Factor (MOM): spread in returns between high prior return portfolios and low prior return portfolios.
- Chicago Board Options Exchange Volatility Index (VIX): measures the implied volatility of SP500 index options. It is often considered the market's "fear gauge" because it intends to represent investors expectation of future 30 days volatility.
- Portfolios formed on Size (PSize): contains return values of portfolios constructed on the market capitalization, namely bottom 30% (Lo30), middle 40% (Mid40), top 30% (Hi30) and quintiles (Lo20, Qnt2, Qnt3, Qnt4, Hi20).
- Portfolios formed on Industries (PInd): returns of ten different industries, namely Consumer Non-Durables (NoDur), Consumer Durables (Durbl), Manufacturing (Manuf), Energy (Enrgy), Business Equipment (HiTec), Telecommunications (Telcm), Shops (Shops), Health (Hlth), Utilities (Utils) and Other (Other).

Thus, we analyze the effect of sentiment and attention on stocks having distinct size (e.g., PSize, SMB), industries (e.g., PInd), momentum (e.g., MOM) and book-to-market ratios (e.g., HML).

Several studies defend that investor sentiment and attention may influence stock market variables. For instance, sentiment may have predictive value for returns (e.g., Sprenger et al., 2014; Bollen et al., 2011; Deng et al., 2011; Sabherwal et al., 2011), volatility (e.g., Antweiler and Frank, 2004; Sabherwal et al., 2011) and trading volume (e.g., Antweiler and Frank, 2004; Tetlock, 2007; Sabherwal et al., 2011). Posting volume on social media (e.g., microblogs and message boards) may also add information for the forecasting of returns (e.g., Wysocki, 1998; Antweiler and Frank, 2004), trading volume (e.g., Wysocki, 1998; Sprenger et al., 2014; Oliveira et al., 2013) and volatility (e.g., Antweiler and Frank, 2004). We have focused on the prediction of these three different stock market variables:

- Daily returns of SP500, RSL, DJIA, NDQ, RMRF, SMB, HML, MOM, PSize and PInd. Returns measure changes in the asset value. We calculated the returns of SP500, RSL, DJIA and NDQ using the total return index (RI datatype) retrieved from Datastream as follows:

$$r_t = \frac{RI_t - RI_{t-1}}{RI_{t-1}} * 100 \quad (46)$$

where RI_t and RI_{t-1} are the total return index values of day t and $t - 1$. The returns of the remaining stocks were directly collected from Professor Kenneth French webpage⁴. All values are in percentage.

- Daily trading volume of SP500 and DJIA collected from Datastream. Trading volume is the number of shares traded in a given period of time (values in thousands).
- Daily volatility is measured using the model free estimate given by the VIX index, and also the realized volatility measure given by a realized kernel for SP500, RSL,

⁴ http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

DJIA and NDQ. Since the realized kernel values are very small, we converted them into annualized realized volatility as (Areal and Taylor, 2002):

$$av_t = \sqrt{rk_t} * \sqrt{252} * 100 \quad (47)$$

where av_t is the annualized realized volatility and rk_t is the realized kernel value. Volatility provides a measures of total risk associated with an investment. VIX data is available at the Chicago Board Options Exchange webpage (<http://www.cboe.com/micro/vix/historical.aspx>) and the realized kernel of the referred indices were collected from Oxford-Man Institute of Quantitative Finance (<http://realized.oxford-man.ox.ac.uk/data/download>).

4.6 MODELS

In this work, we tested five different regression methods (Hastie et al., 2008) to predict stock market variables: MR, NN, SVM, RF and an EA method. The classical MR model assumes a linear relationship between several independent variables and a dependent target. This model is very fast to learn, easy to interpret and has been extensively applied in finance.

NN is a system of interconnected neurons whose functioning is inspired by biological NN. The numeric weights linking the neurons are calibrated during the learning process. In this study we applied the multilayer perceptron having one hidden layer with logistic functions. The output node applies a linear function for regression (e.g., returns, volatility, trading volume). The final output is dependent of the selection of initial weights. To address this problem, we apply an ensemble of NNs and calculate the average of the individual predictions (Hastie et al., 2008). The number of nodes in the hidden layer (H) was the only hyperparameter we had to tune in this work.

The SVM model was initially proposed to perform classification tasks and then extended to regression by adopting an ϵ -insensitive loss function. SVM can execute a

nonlinear mapping by projecting the inputs into high-dimensional space using kernel functions. When compared with multilayer perceptron, the SVM algorithm has the advantage of always converging to the optimal set of weights. In this work, we utilize the popular gaussian kernel and tuned γ , C and ϵ hyperparameters.

RF is an ensemble model that generates a larger number of unpruned decision trees during the training process. The individual trees are based on a random feature selection, using bootstrap training samples. The final RF predictions are built by averaging the outputs of the individual trees.

Ensembles can be used to combine multiple prediction models. In this work, we assume a simple and popular EA approach that often leads to good results and that consists in producing a new predictive response based on the averaging predictions of the MR, NN, SVM and RF models.

Each regression method has its own advantages and disadvantages. MR does not have hyperparameters to tune, is easy to interpret and is less prone to overfitting. However, all other methods (NN, SVM, RF and EA) are more flexible and able to extract nonlinear complex variable associations. Functional nonlinear methods, such as NN and SVM, have various hyperparameters to be tuned (e.g., hidden nodes, kernel parameter) and have more tendency to overfit. SVM has some theoretical advantages over NN, such as the absence of local minima. RF usually obtains good results with its default parameters but it requires higher computational effort for several small or medium sized data sets than a single NN or SVM training. Ensembles often achieve better performances than their individual prediction models (Oztekin et al., 2016). The tested EA is quite simple and does not demand a significant extra computation but requires that all of its individual models are previously trained.

In order to evaluate the relevance of microblogging data to predict stock market behavior, we test all five regression models under two main strategies: with and without microblogging data. The baseline model is the AR model of order p (past time lags) of the predicted target, while the microblog based models are described next. When comparing the baseline and microblog based models, we select the best regression method

(among MR, NN, SVM, RF and EA). Given the extensive number of experiments conducted in this study, we fixed the number of adopted time lags to $p = 5$ for both baseline, i.e., AR(5), and microblog based models. We note that several other related works also used a similar short and fixed number of time lags. For instance, the same five past trading days were used in (Tetlock, 2007; Deng et al., 2011; Oh and Sheng, 2011).

All predictive experiments were conducted using the R tool and rminer package, which facilitates the application of DM methods in real-world tasks (Cortez, 2010). The same methods (MR, NN, SVM and RF) were executed for both baseline and microblog based strategies. The default parameters (e.g., 500 trees for RF, ensemble with 3 multilayer perceptrons for NN, tolerance termination criterion of 0.001 for SVM) were used for the all methods except for the hyperparameters (H for NN and γ , C or ϵ for SVM). These hyperparameters were set using a grid search and internal 10-fold cross-validation considering the first training window (e.g., first 300 trading days for the daily stock market variables). The grid search values were set to $H \in \{1, 2, 3, 4\}$, $\gamma \in \{2^{-9}, 2^{-7}, \dots, 2^3\}$, $C \in \{2^{-3}, 2^{-1}, \dots, 2^9\}$ and $\epsilon \in \{0, 0.01, 0.1, 0.2, 0.4, 0.6\}$.

The tested models to predict returns were:

$$\begin{aligned}
 \hat{R}_t &= f(R_{t-1}, R_{t-2}, R_{t-3}, R_{t-4}, R_{t-5}) & \text{(MRet1, baseline)} \\
 \hat{R}_t &= f(S_{t-1}, S_{t-2}, S_{t-3}, S_{t-4}, S_{t-5}) & \text{(MRet2)} \\
 \hat{R}_t &= f(R_{t-1}, R_{t-2}, R_{t-3}, R_{t-4}, R_{t-5}, S_{t-1}, S_{t-2}, S_{t-3}, S_{t-4}, S_{t-5}) & \text{(MRet3)} \\
 \hat{R}_t &= f(S_{t-1}, S_{t-2}, S_{t-3}, S_{t-4}, S_{t-5}, Nt_{t-1}, Nt_{t-2}, Nt_{t-3}, Nt_{t-4}, Nt_{t-5}) & \text{(MRet4)} \\
 \hat{R}_t &= f(R_{t-1}, R_{t-2}, R_{t-3}, R_{t-4}, R_{t-5}, S_{t-1}, S_{t-2}, S_{t-3}, S_{t-4}, S_{t-5}, \\
 & Nt_{t-1}, Nt_{t-2}, Nt_{t-3}, Nt_{t-4}, Nt_{t-5}) & \text{(MRet5)} \\
 \hat{R}_t &= f(Nt_{t-1}, Nt_{t-2}, Nt_{t-3}, Nt_{t-4}, Nt_{t-5}) & \text{(MRet6)} \\
 \hat{R}_t &= f(R_{t-1}, R_{t-2}, R_{t-3}, R_{t-4}, R_{t-5}, Nt_{t-1}, Nt_{t-2}, Nt_{t-3}, Nt_{t-4}, Nt_{t-5}) & \text{(MRet7)}
 \end{aligned} \tag{48}$$

where S_t refers to the sentiment value of day t (VA, BI, BullR or KF indicators) and Nt_t refers to the first difference of posting volume of day t .

To forecast volatility, we experimented the following models:

$$\begin{aligned}
 \hat{V}_t &= f(V_{t-1}, V_{t-2}, V_{t-3}, V_{t-4}, V_{t-5}) && \text{(MVt1, baseline)} \\
 \hat{V}_t &= f(S_{t-1}, S_{t-2}, S_{t-3}, S_{t-4}, S_{t-5}) && \text{(MVt2)} \\
 \hat{V}_t &= f(V_{t-1}, V_{t-2}, V_{t-3}, V_{t-4}, V_{t-5}, S_{t-1}, S_{t-2}, S_{t-3}, S_{t-4}, S_{t-5}) && \text{(MVt3)} \\
 \hat{V}_t &= f(S_{t-1}, S_{t-2}, S_{t-3}, S_{t-4}, S_{t-5}, N_{t-1}, N_{t-2}, N_{t-3}, N_{t-4}, N_{t-5}) && \text{(MVt4)} \\
 \hat{V}_t &= f(V_{t-1}, V_{t-2}, V_{t-3}, V_{t-4}, V_{t-5}, S_{t-1}, S_{t-2}, S_{t-3}, S_{t-4}, S_{t-5}, \\
 &N_{t-1}, N_{t-2}, N_{t-3}, N_{t-4}, N_{t-5}) && \text{(MVt5)} \\
 \hat{V}_t &= f(N_{t-1}, N_{t-2}, N_{t-3}, N_{t-4}, N_{t-5}) && \text{(MVt6)} \\
 \hat{V}_t &= f(V_{t-1}, V_{t-2}, V_{t-3}, V_{t-4}, V_{t-5}, N_{t-1}, N_{t-2}, N_{t-3}, N_{t-4}, N_{t-5}) && \text{(MVt7)} \\
 &&& (49)
 \end{aligned}$$

where V_t is the volatility value of day t .

In the prediction of trading volume, we tested the following models:

$$\begin{aligned}
 \hat{V}_0 &= f(V_{0t-1}, V_{0t-2}, V_{0t-3}, V_{0t-4}, V_{0t-5}) && \text{(MVol1, baseline)} \\
 \hat{V}_0 &= f(S_{t-1}, S_{t-2}, S_{t-3}, S_{t-4}, S_{t-5}) && \text{(MVol2)} \\
 \hat{V}_0 &= f(V_{0t-1}, V_{0t-2}, V_{0t-3}, V_{0t-4}, V_{0t-5}, S_{t-1}, S_{t-2}, S_{t-3}, S_{t-4}, S_{t-5}) && \text{(MVol3)} \\
 \hat{V}_0 &= f(S_{t-1}, S_{t-2}, S_{t-3}, S_{t-4}, S_{t-5}, N_{t-1}, N_{t-2}, N_{t-3}, N_{t-4}, N_{t-5}) && \text{(MVol4)} \\
 \hat{V}_0 &= f(V_{0t-1}, V_{0t-2}, V_{0t-3}, V_{0t-4}, V_{0t-5}, S_{t-1}, S_{t-2}, S_{t-3}, S_{t-4}, S_{t-5}, \\
 &N_{t-1}, N_{t-2}, N_{t-3}, N_{t-4}, N_{t-5}) && \text{(MVol5)} \\
 \hat{V}_0 &= f(N_{t-1}, N_{t-2}, N_{t-3}, N_{t-4}, N_{t-5}) && \text{(MVol6)} \\
 \hat{V}_0 &= f(V_{0t-1}, V_{0t-2}, V_{0t-3}, V_{0t-4}, V_{0t-5}, N_{t-1}, N_{t-2}, N_{t-3}, N_{t-4}, N_{t-5}) && \text{(MVol7)} \\
 &&& (50)
 \end{aligned}$$

where V_0 is the trading volume of day t .

In this study, we also used Twitter and KF sentiment indicators to predict survey sentiment indicators. AAI and II are popular survey sentiment indicators already used in diverse research studies about stock market behavior (e.g., Solt and Statman, 1988; Fisher and Statman, 2000; Verma and Soydemir, 2009). These indicators have distinct publication days. AAI values are released each Thursday comprising data from previous Thursday until Wednesday. II indicators are published each Wednesday

and they are related to analysis performed from previous Wednesday to last Tuesday. To properly compare Twitter and KF indicators to each survey measure, we created different weekly microblogging indicators corresponding to the time interval used by each survey indicator. Additionally, we tested the utilization of the previous 7 daily (one week) Twitter and KF indicators.

We predicted four different values for each survey indicator: VA, BI, negative and positive percentage. In the forecasting of each survey value, we used the equivalent Twitter indicator (i.e., VA, BI, BearR and BullR). To compute the weekly Twitter values, we also experimented two different approaches. The first approach (Aggregate Approach (AA)) calculates the weekly value using the total number of positive and/or negative messages of the week while the second approach (Mean Approach (MA)) computes the average of the seven daily indicators that compose the week. The production of the weekly KF values applies the second approach.

We applied seven different models by exploring five lags of the target survey indicator, five lags of the weekly microblogging indicators and seven lags (one week) of the daily microblogging indicators. The explored predictive models of survey sentiment are:

$$\begin{aligned}
 \hat{Sv}_t &= f(Sv_{t-1}, Sv_{t-2}, Sv_{t-3}, Sv_{t-4}, Sv_{t-5}) && \text{(MSv1, baseline)} \\
 \hat{Sv}_t &= f(Sw_{t-1}, Sw_{t-2}, Sw_{t-3}, Sw_{t-4}, Sw_{t-5}) && \text{(MSv2)} \\
 \hat{Sv}_t &= f(Sd_{t-1}, Sd_{t-2}, Sd_{t-3}, Sd_{t-4}, Sd_{t-5}, Sd_{t-6}, Sd_{t-7}) && \text{(MSv3)} \\
 \hat{Sv}_t &= f(Sw_{t-1}, Sw_{t-2}, Sw_{t-3}, Sw_{t-4}, Sw_{t-5}, \\
 &Sd_{t-1}, Sd_{t-2}, Sd_{t-3}, Sd_{t-4}, Sd_{t-5}, Sd_{t-6}, Sd_{t-7}) && \text{(MSv4)} \\
 \hat{Sv}_t &= f(Sv_{t-1}, Sv_{t-2}, Sv_{t-3}, Sv_{t-4}, Sv_{t-5}, \\
 &Sw_{t-1}, Sw_{t-2}, Sw_{t-3}, Sw_{t-4}, Sw_{t-5}) && \text{(MSv5)} \\
 \hat{Sv}_t &= f(Sv_{t-1}, Sv_{t-2}, Sv_{t-3}, Sv_{t-4}, Sv_{t-5}, \\
 &Sd_{t-1}, Sd_{t-2}, Sd_{t-3}, Sd_{t-4}, Sd_{t-5}, Sd_{t-6}, Sd_{t-7}) && \text{(MSv6)} \\
 \hat{Sv}_t &= f(Sv_{t-1}, Sv_{t-2}, Sv_{t-3}, Sv_{t-4}, Sv_{t-5}, \\
 &Sw_{t-1}, Sw_{t-2}, Sw_{t-3}, Sw_{t-4}, Sw_{t-5}, \\
 &Sd_{t-1}, Sd_{t-2}, Sd_{t-3}, Sd_{t-4}, Sd_{t-5}, Sd_{t-6}, Sd_{t-7}) && \text{(MSv7)}
 \end{aligned} \tag{51}$$

where Sv_t corresponds to the weekly survey sentiment values (AAII or II) of day t , Sw_t corresponds to the weekly microblogging sentiment values (BI, VA, BullR, BearR or KF) of day t and Sd_t corresponds to the daily microblogging values (AAII or II) of day t .

4.7 EVALUATION

There is empirical evidence that good forecasting methods provide consistent results across multiple metrics (Crone et al., 2011). Given the large number of experiments conducted in this work, we opted for a single error metric when evaluating the quality of the predictions. In this work, we selected an absolute error based metric, which is a common approach in the forecasting domain (Hyndman and Koehler, 2006). For instance, in (Armstrong and Collopy, 1992; Armstrong, 2001) it is argued that squared error metrics, such as Root Mean Square Error, are not reliable due to their sensitivity

to outliers and should be replaced by absolute error metrics when comparing across time series. We note that other related works also have adopted absolute error metrics, such as: Mean Absolute Error (MAE) (Deng et al., 2011) and Mean Absolute Percentage Error (MAPE) (Bollen et al., 2011; Deng et al., 2011; Mao et al., 2011).

Using any absolute error measure should lead to the same ranking differences when comparing distinct forecasting models, thus the particular choice of such measure affects mostly its range of values and interpretation. In this study, we selected the Normalized Mean Absolute Error (NMAE) metric that is calculated as (Goldberg et al., 2001):

$$\begin{aligned} MAE &= \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \\ NMAE &= \frac{MAE}{y_H - y_L} \end{aligned} \quad (52)$$

where y_i is the target value for time i , y_H is the highest target value, y_L is the lowest target value, \hat{y}_i is the predicted value and N corresponds to the number of predictions considered. When compared with other absolute based metrics, the NMAE presents several advantages. First, it is easier to interpret than MAE, since it expresses the error as a percentage of the full target scale. The lower the NMAE values, the better are the forecasts. Second, it is scale independent, which is particularly useful in this work since we predict variables with distinct scales. Third, it does not contain the limitations of other scale independent measures. For instance, MAPE can produce infinity values when the denominator (target values) is zero (Hyndman and Koehler, 2006), which might occur in several of the predicted variables (e.g., returns can have near zero values). While we only consider NMAE values, in the tables with predictive results we also show the target range value ($y_H - y_L$), thus the MAE values can be easily obtained by computing the inverse of Equation 52.

To measure the generalization capability of the predictive models, we applied a fixed-size rolling windows scheme (Tashman, 2000). For each prediction (e.g., day t), the model is trained using a window of the previous W consecutive samples (e.g., from day $t - W$ to day $t - 1$) and used to predict the next value (time t). Then, the training

window is slid by discarding its oldest element and adding the value of t in order to retrain the model and predict the value at time $t + 1$, and so on. Therefore, a data set of length L will produce $L - W$ model trainings and their respective predictions. This rolling windows validation is realistic since it mimics the way a predictive model would be used in a real-environment, trained with a large number of past data and used to predict the next daily/weekly values. And it is robust, since it allows the training and testing of a large number of models. In this work, we applied a window size of $W=300$ days for the prediction of the daily stock market variables and $W=50$ observations for the forecasting of the weekly survey sentiment indicators. The number of predictions range from 392 to 439 for the daily variables and 92 to 93 for the weekly survey indices. We note that these numbers are much higher than most state of the art works (e.g., 8 and 30 predictions).

The prediction ability of the models was evaluated by the DMST test for predictive accuracy (Diebold and Mariano, 1995), under a pairwise comparison between the baseline and microblog based models. Thus, we assume that the microblogging data has predictive content if the respective model has a statistical significant DMST test.

4.8 EXPERIMENTAL RESULTS AND DISCUSSION

4.8.1 *Prediction of Returns*

This work analyzed the prediction of returns of diverse indices as well as portfolios formed on size and industries. We tested sentiment indicators created by three different formulas (BullR, BI and VA) and KF procedure. Table 12 presents the NMAE results produced by the four regression methods (MR, RF, NN and SVM) using sentiment indicators produced by BullR approach and the p-value calculated by DMST test for predictive accuracy (Diebold and Mariano, 1995) for SP500, RSL, DJIA, NDQ, RMRF, SMB, HML and MOM. In the table, the baseline results are underlined, while

the lowest NMAE value is in **bold**. The first column also shows in brackets the number of predictions ($L - W$) and the target range ($y_H - y_L$).

Table 12.: Predictive results for returns of SP500, RSL, DJIA, NDQ, RMRF, SMB, HML and MOM using general sentiment indicators calculated by BullR formula and first difference of posting volume

	Mtd	MRet ₁	MRet ₂	MRet ₃	MRet ₄	MRet ₅	MRet ₆	MRet ₇
(number of predictions: 414; returns range: 7.53)	DJIA MR	<u>8.12</u>	8.11	8.24	8.20	8.33	8.11	8.25
	RF	8.38	8.41	8.43	8.44	8.42	8.38	8.45
	SVM	8.19	8.09	8.01*	8.07	8.07	8.05	8.05
	NN	8.28	8.42	8.16	8.29	8.44	8.56	8.26
	EA	8.14	8.2	8.16	8.15	8.23	8.13	8.22
(number of predictions: 392; returns range: 3.36)	HML MR	10.49	10.37	10.54	10.36	10.54	10.34	10.52
	RF	10.42	10.56	10.39	10.49	10.49	10.62	10.56
	SVM	10.29	10.31	10.24	10.26	10.25	10.26	10.33
	NN	10.92	10.84	10.66	10.62	10.73	10.99	10.47
	EA	10.55	10.32	10.37	10.33	10.37	10.35	10.4
(number of predictions: 392; returns range: 4.63)	MOM MR	10.91	10.80	10.94	11.02	11.18	11.03	11.11
	RF	11.00	11.14	10.93	11.18	11.17	11.31	11.21
	SVM	10.78	10.79	10.73	10.84	10.78	10.88	10.81
	NN	11.18	11.01	11.99	11.33	11.32	11.13	11.11
	EA	10.8	10.84	10.92	11.31	11.1	11.08	11.01
(number of predictions: 439; returns range: 9.35)	NDQ MR	7.68	7.70	7.82	7.86	7.99	7.74	7.85
	RF	7.85	7.90	7.88	7.97	7.96	7.88	7.93
	SVM	<u>7.61</u>	7.59	7.61	7.78	7.62	7.64	7.69
	NN	7.99	8.38	7.94	8.93	7.81	7.97	8.04
	EA	7.71	7.74	7.7	7.81	7.85	7.7	7.88
(number of predictions: 392; returns range: 7.58)	RMRF MR	8.31	8.32	8.45	8.42	8.58	8.36	8.44
	RF	8.54	8.59	8.61	8.68	8.81	8.56	8.74
	SVM	8.27	8.35	8.28	8.24	8.24	8.24	8.26
	NN	8.95	8.46	8.47	8.65	8.54	9.57	8.61
	EA	8.36	8.38	8.38	8.43	8.41	8.35	8.53
(number of predictions: 439; returns range: 7.02)	RSL MR	11.10	11.31	11.35	11.46	11.49	11.28	11.30
	RF	11.17	11.47	11.19	11.58	11.40	11.44	11.33
	SVM	11.15	11.71	11.20	11.24	11.19	11.13	11.13
	NN	11.21	11.60	11.21	11.46	11.64	12.57	13.85
	EA	11.02	11.37	11.44	11.36	11.32	11.34	11.3
(number of predictions: 392; returns range: 3.36)	SMB MR	12.44	12.44	12.59	12.50	12.76	12.40	12.59
	RF	12.46	12.72	12.54	12.74	12.50	12.72	12.42
	SVM	12.44	12.41	12.40	12.27*	12.46	12.48	12.44
	NN	13.49	13.09	13.05	13.41	12.70	12.55	12.89
	EA	12.5	12.66	12.48	12.4	12.46	12.44	12.4
(number of predictions: 413; returns range: 7.85)	SP500 MR	7.90	7.88	8.01	7.97	8.11	7.92	8.01
	RF	8.32	8.20	8.32	8.23	8.35	8.13	8.32
	SVM	7.87	7.83	7.88	7.92	7.86	7.83	7.80**
	NN	8.59	8.17	7.95	8.00	8.02	7.92	8.40
	EA	7.92	7.99	7.94	8	7.99	8.11	7.97

For each index, the baseline model is underlined and the lowest NMAE value is in **bold** (* – p-value < 10%, ** – p-value < 5%, *** – p-value < 1%, NMAE in %)

We summarize the results for BullR, BI, VA and KF indicators in Table 13. For each index, Table 13 identifies the baseline model, the lowest NMAE model and models generating statistically significant results in the DMST test. Ten SVM models significantly improve the results of baseline models for the forecasting of SP500, DJIA, MOM, SMB and RMRF. Two of these models obtain p-value inferior to 5%, both for the forecasting of SP500. There are no microblogging models having significantly higher predictive

accuracy than baseline models for the remaining indices (i.e., HML, NDQ and RSL). However, the lowest NMAE values are produced by models containing microblogging features for all items, except for RSL. The utilization of KF sentiment indicators lowered the NMAE results for four indices: DJIA, MOM, RMRF and SP500. Furthermore, models applying KF indicators are significantly more accurate than baseline for SP500, DJIA, MOM and RMRF, while models using TWT indicators obtain this statistical significant results only for DJIA, SMB and SP500. Moreover, SVM MRet3 model using KF indicators produces statistically better forecasts of the SP500 than the baseline. The posting volume features were important for some indices. The lowest NMAE values of the prediction of NDQ and SMB were produced by models using the first difference of the number of tweets. Additionally, there are diverse models applying these features significantly more accurate than baseline for SP500 and SMB. The majority of the most accurate models applies the SVM method.

Table 13.: Predictive results for returns of SP500, RSL, DJIA, NDQ, RMRF, SMB, HML and MOM using general sentiment indicators (BullR, BI and VA approaches), KF indicators and first difference of the posting volume

Index	Baseline	Lowest NMAE	Statistical significant results
DJIA (n predictions: 414; returns range: 7.53)	MR: 8.12	SVM MRet2 (KF): 7.98*	SVM MRet3 (BullR): 8.01* SVM MRet2 (KF): 7.98*
HML (n predictions: 392; returns range: 3.36)	SVM: 10.29	SVM MRet3 (BullR): 10.24	
MOM (n predictions: 392; returns range: 4.63)	SVM: 10.78	SVM MRet2 (KF): 10.69*	SVM MRet2 (KF): 10.69*
NDQ (n predictions: 439; returns range: 9.35)	SVM: 7.61	SVM MRet7: 7.58	
RMRF (n predictions: 392; returns range: 7.58)	SVM: 8.27	SVM MRet3 (KF): 8.19*	SVM MRet3 (KF): 8.19*
RSL (n predictions: 439; returns range: 7.02)	EA: 11.02	EA MRet1: 11.02	
SMB (n predictions: 392; returns range: 3.36)	MR: 12.44 SVM: 12.44	SVM MRet4 (BullR): 12.27*	SVM MRet4 (BullR): 12.27*
SP500 (n predictions: 439; returns range: 7.85)	SVM: 7.87	SVM MRet3 (KF): 7.79**	SVM MRet7: 7.80** SVM MRet6: 7.81* SVM MRet4 (VA): 7.81* SVM MRet5 (VA): 7.81* SVM MRet3 (KF): 7.79**

For each index: NMAE values of baseline model, lowest NMAE model and models producing statistical significant results in DMST test (* – p-value < 10%, ** – p-value < 5%, *** – p-value < 1%, sentiment indicators in parenthesis, NMAE in %, NMAE values lower than baseline are in **bold**)

Some studies refer that there is a distinct sentiment effect in prices of stocks with large and small market capitalization (e.g., Neal and Wheatley, 1998; Baker and Wurgler, 2006; Baker et al., 2012; Lee et al., 2002). To analyze the influence of sentiment on

stocks with different size, we predicted returns of portfolios formed on size. Table 14 shows a summary of all tested models for the prediction of portfolios formed on size. There are sixteen models significantly more accurate than baseline in the forecasting of portfolios of lower market capitalization (i.e., Lo20 and Lo30). The prediction of Lo20 returns has one model obtaining p-value less than 1% in the pairwise DMST test and five other models generating p-value inferior to 5%. These results may indicate that sentiment is more informative to the prediction of stocks of smaller capitalization, which is consistent with previous findings (e.g., Baker and Wurgler, 2006; Baker et al., 2012). The most accurate microblogging models are SVM. Posting volume seems informative for the prediction of these portfolios. The majority of the models obtaining statistical significant results in the pairwise DMST test and the lowest NMAE values apply the first difference of posting volume. KF indicators seem less useful for the forecasting of portfolios formed on size than for the prediction of indices. Models applying KF indicators outperformed models using Twitter indicators only for Qnt4. Nevertheless, in the prediction of Lo20 there is one model applying KF indicators that produced p-value less than 5% in the DMST test and another model obtained p-value less than 10%.

We also tested the prediction of portfolios of 10 different industries. The respective evaluation results are summarized in Table 15. There are several microblogging models significantly more accurate than baseline for Enrgy, HiTec and Other. Seven of these models obtain p-value less than 5% in the pairwise DMST test for the prediction of Enrgy and three models have p-value less than 5% for the forecasting of HiTec. Therefore, microblogging features are particularly informative for Enrgy and HiTec. A possible explanation for these results is the unusual high number of microblogging messages related to stocks of these sectors, mainly for HiTec. Thus, the general sentiment indicators may be biased toward these industries. For demonstration purposes, the forecasted returns of Enrgy sector by the NN MRet2 (VA) are presented in the Figure 20. The utilization of posting volume was also important. The majority of models obtaining statistical significant results in the pairwise DMST test use posting

Chapter 4. The Impact of Microblogging Data for Stock Market Prediction

Table 14.: Predictive results for returns of portfolios formed on size using general sentiment indicators (BullR, BI and VA approaches), KF indicators and first difference of the posting volume

Index	Baseline	Lowest NMAE	Statistical significant results
Lo30 (returns range: 5.96)	MR: 11.75	SVM MRet4 (VA): 11.57*	SVM MRet4 (BullR): 11.59* SVM MRet4 (BI): 11.59* SVM MRet4 (VA): 11.57* SVM MRet5 (VA): 11.60*
Med40 (returns range: 6.87)	EA: 11.04	SVM MRet4 (BullR): 11.01 SVM MRet5 (BullR): 11.01 SVM MRet7: 11.01 SVM MRet5 (BI): 11.01 SVM MRet5 (KF): 11.01	
Hi30 (returns range: 7.34)	SVM: 8.71	SVM MRet7 (BI): 8.68 SVM MRet4 (VA): 8.68 SVM MRet5 (VA): 8.68 SVM MRet6: 8.68	
Lo20 (returns range: 5.64)	MR: 11.94 EA: 11.94	SVM MRet4 (VA): 11.66***	SVM MRet4 (BullR): 11.74* SVM MRet4 (BI): 11.75* SVM MRet5 (BI): 11.70** SVM MRet7: 11.75** SVM MRet2 (VA): 11.75* EA MRet2 (VA): 11.75* SVM MRet3 (VA): 11.78* SVM MRet4 (VA): 11.66*** SVM MRet5 (VA): 11.74** SVM MRet6: 11.71** SVM MRet2 (KF): 11.74* SVM MRet4 (KF): 11.72**
Qnt2 (returns range: 7.31)	EA: 11.56	SVM MRet6: 11.53	
Qnt3 (returns range: 6.94)	MR: 10.83 EA: 10.83	SVM MRet5 (BullR): 10.77	
Qnt4 (returns range: 6.72)	SVM: 10.31	SVM MRet2 (KF): 10.28	
Hi20 (returns range: 7.33)	SVM: 8.63	SVM MRet6: 8.57 SVM MRet7: 8.57	

For each portfolio: NMAE values of baseline model, lowest NMAE model and models producing statistical significant results in DMST test (* – p-value < 10%, ** – p-value < 5%, *** – p-value < 1%, sentiment indicators in parenthesis, NMAE in %, NMAE values lower than baseline are in **bold**, 392 predictions)

volume features. Moreover, the utilization of KF indicators is beneficial in some situations. KF indicators permit lower NMAE values than TWT indicators for the HiTec and Utils sectors, and they are the most applied sentiment indicator in models which are significantly more accurate than baselines.

The application of sectorial sentiment indicators may allow a better analysis of the effect of sentiment in each sector. We had to exclude the “Other” sector from this analysis because there was no information about its SIC codes, so we were unable to create its sectorial indicator. Additionally, there are no tweets for Durbl sector for 37 days, so we excluded those days from the analysis. Table 16 present the results for these models.

Sectorial indicators are particularly useful for Shops, HiTec and Telcm industries. There are more models significantly more accurate than baseline using sectorial indi-

Table 15.: Predictive results for returns of portfolios formed on industries using general sentiment indicators (BullR, BI and VA approaches), KF indicators and first difference of the posting volume

Index	Baseline	Lowest NMAE	Statistical significant results
Durbl (returns range: 5.71)	MR: 13.90	EA MRet2 (BullR): 13.74 SVM MRet6: 13.74	
Enrgy (returns range: 21.12)	SVM: 7.18	NN MRet2 (VA): 7.07**	SVM MRet3 (BI): 7.10** SVM MRet5 (BI): 7.10** SVM MRet6: 7.13* SVM MRet7: 7.12* SVM MRet2 (VA): 7.08** NN MRet2 (VA): 7.07** SVM MRet3 (VA): 7.11* SVM MRet5 (VA): 7.10** SVM MRet2 (KF): 7.10** SVM MRet3 (KF): 7.10** SVM MRet4 (KF): 7.11* SVM MRet5 (KF): 7.12*
HiTec (returns range: 5.90)	MR: 12.79 EA: 12.79	SVM MRet5 (KF): 12.55**	SVM MRet3 (BullR): 12.64* SVM MRet4 (BullR): 12.56** SVM MRet4 (BI): 12.56** SVM MRet2 (KF): 12.58* SVM MRet3 (KF): 12.64* SVM MRet5 (KF): 12.55**
Hlth (returns range: 7.10)	NN: 12.94	SVM MRet6: 12.85	
Manuf (returns range: 5.17)	SVM: 13.66	SVM MRet7: 13.62	
NoDur (returns range: 4.55)	SVM: 12.77	SVM MRet3 (BullR): 12.72	
Other (returns range: 4.04)	SVM: 13.51	SVM MRet6: 13.34* SVM MRet4 (VA): 13.34* SVM MRet4 (KF): 13.34*	SVM MRet6: 13.34* SVM MRet3 (VA): 13.37* SVM MRet4 (VA): 13.34* SVM MRet4 (KF): 13.34*
Shops (returns range: 4.76)	SVM: 13.85	SVM MRet4 (BI): 13.67*	SVM MRet4 (BI): 13.67*
Telcm (returns range: 5.52)	SVM: 14.15	SVM MRet2 (BullR): 13.98*	SVM MRet2 (BullR): 13.98*
Utils (returns range: 5.77)	SVM: 11.18	RF MRet5 (KF): 11.04	

For each portfolio: NMAE values of baseline model, lowest NMAE model and models producing statistical significant results in DMST test (* – p-value < 10%, ** – p-value < 5%, *** – p-value < 1%, sentiment indicators in parenthesis, NMAE in %, NMAE values lower than baseline are in **bold**, 350 predictions)

cators than applying general indicators. The SVM model using previous values of returns, posting volume and BI sentiment (SVM MRet5 (BI)) obtains a p-value less than 5% for the prediction of Telcm. Moreover, the lowest NMAE values for Telcm and Shops are obtained by models utilizing sectorial indicators.

In summary, microblogging sentiment and attention indicators were particularly informative for the forecasting of returns of SP500, portfolios of smaller market capitalization (Lo20 and Lo30) and some sectors such as HiTec, Enrgy and Telcm. In the prediction of the returns of the mentioned stocks, there are diverse microblogging models obtaining p-value less than 5% in the pairwise DMST test. Many of these models apply both sentiment and posting volume indicators. These results may suggest that sentiment and attention have more impact on future returns of stocks of inferior capitalization and technology related companies. The impact of sentiment on these type

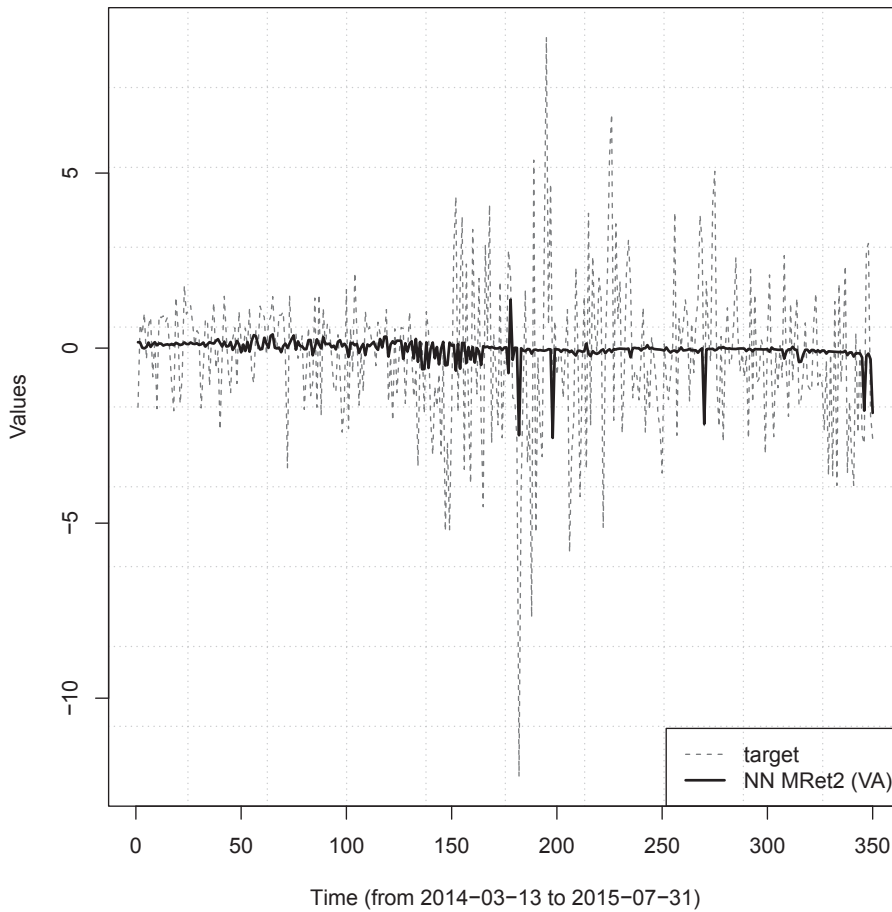


Figure 20.: Predicted and real values for Enrgy

of stocks may be explained by a higher concentration of irrational investors. Small stocks are considered to be mostly held by individual investors and less attractive to rational investors (Baker and Wurgler, 2006; Baker et al., 2012). Additionally, a considerable part of technology companies are young and have a small track record. So, they are less appealing to professional investors that prefer easier to value stocks (Baker and Wurgler, 2006). The unpredictability of irrational traders adds risk on prices and makes arbitrage strategies more difficult to implement (Long et al., 1990). Thus, stock prices may differ from fundamental values because arbitrage may be insufficient to instantly eliminate mispricing generated by investor sentiment. In these situations, there is margin to predict future prices and sentiment indicators can be informative.

Table 16.: Predictive results for returns of portfolios formed on industries using sectorial sentiment indicators (BullR, BI and VA approaches) and first difference of the posting volume

Index	Baseline	Lowest NMAE	Statistical significant results
Durbl (returns range: 5.71)	MR: 13.90	EA MRet2 (BI): 13.70	
Enrgy (returns range: 21.12)	SVM: 7.18	SVM MRet3 (BullR): 7.10* SVM MRet3 (BI): 7.10** SVM MRet5 (BI): 7.10**	SVM MRet3 (BullR): 7.10* SVM MRet4 (BullR): 7.11* SVM MRet3 (BI): 7.10** SVM MRet5 (BI): 7.10**
HiTec (returns range: 5.90)	MR: 12.79 EA: 12.79	EA MRet2 (BullR): 12.55**	SVM MRet2 (BullR): 12.57** SVM MRet4 (BullR): 12.58* EA MRet2 (BullR): 12.55** SVM MRet6: 12.56** SVM MRet2 (BI): 12.61* EA MRet2 (BI): 12.59* SVM MRet4 (BI): 12.62* SVM MRet4 (VA): 12.56**
Hlth (returns range: 7.10)	NN: 12.94	NN MRet1: 12.94	
Manuf (returns range: 5.17)	SVM: 13.66	SVM MRet5 (VA): 13.60	
NoDur (returns range: 4.55)	SVM: 12.77	MR MRet2 (VA): 12.72	
Shops (returns range: 4.76)	SVM: 13.85	SVM MRet5 (BI): 13.63*	SVM MRet5 (BullR): 13.64* SVM MRet6: 13.67* SVM MRet5 (BI): 13.63* SVM MRet4 (VA): 13.65*
Telcm (returns range: 5.52)	SVM: 14.15	SVM MRet4 (BullR): 13.92*	SVM MRet4 (BullR): 13.92* SVM MRet5 (BullR): 13.95* SVM MRet5 (BI): 13.93**
Utils (returns range: 5.77)	SVM: 11.18	SVM MRet6: 11.14	

For each portfolio: NMAE values of baseline model, lowest NMAE model and models producing statistical significant results in DMST test (* – p-value < 10%, ** – p-value < 5%, *** – p-value < 1%, sentiment indicators in parenthesis, NMAE in %, NMAE values lower than baseline are in **bold**, 350 predictions)

The utilization of KF indicators was important in some situations. For instance, models applying KF indicators have p-value less than 5% for the prediction of SP500, Lo20, HiTec and Enrgy. Moreover, the usage of KF sentiment indicators decreased the NMAE values for DJIA, MOM, RMRF, SP500, Qnt4, HiTec and Utils. The sectorial indicators were useful for the prediction of returns of Shops and Telcm industries.

Regarding the tested methods, SVM was clearly the most effective. Therefore, the relationship between sentiment, attention and returns may be nonlinear.

4.8.2 Prediction of Volatility

In this study, we forecasted VIX and the annualized realized volatility of SP500, RSL, DJIA and NDQ. Table 17 outlines the results produced by models applying posting volume and general sentiment indicators computed by BullR, BI, VA, AG and KF approaches.

Table 17.: Predictive results for VIX and annualized realized volatility of SP500, RSL, DJIA and NDQ using general sentiment indicators (BullR, BI, VA and AG approaches), KF indicators and first difference of the posting volume

Index	Baseline	Lowest NMAE	Statistical significant results
DJIA (n predictions: 413; realized volatility range: 92.41)	MR: 2.91	SVM MVlt ₃ (AG): 2.79	MR MVlt ₃ (KF): 2.85*
NDQ (n predictions: 413; realized volatility range: 57.18)	MR: 3.89	SVM MVlt ₃ (VA): 3.87	
RSL (n predictions: 412; realized volatility range: 39.56)	MR: 5.71	MR MVlt ₁ : 5.71	
SP500 (n predictions: 413; realized volatility range: 67.90)	EA: 3.34	EA MVlt ₁ : 3.34 EA MVlt ₃ (AG): 3.34	
VIX (n predictions: 413; VIX range: 30.42)	EA: 3.26	SVM MVlt ₃ (BullR): 3.25	

For each index: NMAE values of baseline model, lowest NMAE model and models producing statistical significant results in DMST test (* – p-value < 10%, ** – p-value < 5%, *** – p-value < 1%, sentiment indicators in parenthesis, NMAE in %, NMAE values lower than baseline are in **bold**)

The utilization of KF sentiment values and previous volatility values significantly improves the forecasting of DJIA realized volatility compared to the AR(5) model. However, the lowest p-value of DMST test is not inferior to 5%. AG indicators were applied in the model that obtained the lowest NMAE value for DJIA and SP500. However, this model is not significantly more accurate than baseline. The inclusion of the number of tweets do not seems to benefit the forecasting of volatility. There are no models utilizing posting volume that significantly outperform the respective baseline models. Also, the most accurate models using posting volume features do not improve the results of models without posting volume information. MR, EA and SVM are the most effective methods.

4.8.3 Prediction of Trading Volume

In this study, we predicted the trading volume of SP500 and DJIA. Table 18 digests the results for models using general sentiment indicators, KF indicators and first difference of the number of tweets.

The application of microblogging sentiment values and the first difference of the number of tweets significantly improves baseline results for the prediction of DJIA trading volume in two situations: SVM MVlt₅ (BullR) and SVM MVlt₅ (BI). However,

Table 18.: Predictive results for trading volume of SP500 and DJIA using general sentiment indicators (BullR, BI, VA and AG approaches), KF indicators and first difference of the posting volume

Index	Baseline	Lowest NMAE	Statistical significant results
DJIA (n predictions: 414; volume range: 310804)	SVM: 6.00	SVM MVLt ₅ (BullR): 5.84* SVM MVLt ₅ (BI): 5.85*	SVM MVLt ₅ (BullR): 5.84*
SP500 (n predictions: 413; volume range: 1636036)	SVM: 4.98	SVM MVLt ₁ : 4.98	

For each index: NMAE values of baseline model, lowest NMAE model and models producing statistical significant results in DMST test (* – p-value < 10%, ** – p-value < 5%, *** – p-value < 1%, sentiment indicators in parenthesis, NMAE in %, NMAE values lower than baseline are in **bold**)

the lowest NMAE values for SP500 are obtained by the AR(5) model. These results add some evidence that posting volume and sentiment may have predictive content for forecasting of trading volume (e.g., Sabherwal et al., 2011; Antweiler and Frank, 2004; Tetlock, 2007; Wysocki, 1998; Oliveira et al., 2013; Sprenger et al., 2014). The utilization of AG indicators do not result in statistical significant DMST tests. Therefore, we do not have evidence that disagreement is associated with trading volume as found in some studies (e.g., Antweiler and Frank, 2004; Shalen, 1993). The most accurate models also did not applied KF indicators. The SVM method is clearly the best performing approach for the forecasting of trading volume.

4.8.4 Prediction of Survey Sentiment Indicators Using Twitter and KF Sentiment Indicators

In this subsection, we apply Twitter and KF indicators to forecast AAI and II indicators. The applied procedure is similar to the applied in the forecasting of stock market variables. However, we reduced the size of the rolling windows to 50 because the weekly periodicity downsized the data set. We tested the prediction of four different survey values: VA, BI, negative and positive values. Table 19 presents the evaluation results of the forecasting of AAI and II indicators using weekly Twitter indicators produced by AA approach (aggregating positive and negative messages of the week). For each prediction, the table also shows the number of predictions and target range ($y_L - y_H$).

Table 19.: Prediction of weekly AAll and II values using Twitter sentiment indicators calculated by AA approach

Mtd	MSv1	MSv2	MSv3	MSv4	MSv5	MSv6	MSv7	Mtd	MSv1	MSv2	MSv3	MSv4	MSv5	MSv6	MSv7
Panel A: Prediction of AAll values calculated by BI formula (93 predictions; range: 58.21)															
MR	14.35	18.10	17.72	19.42	15.52	16.19	16.98	RF	14.61	17.59	17.92	16.95	14.71	15.18	14.69
SVM	14.30	16.92	16.84	16.56	14.41	16.42	14.73	NN	16.11	21.58	21.71	20.61	17.15	17.81	17.39
EA	14.79	17.34	17.41	17.68	14.99	15.77	15.4								
Panel B: Prediction of AAll values calculated by VA formula (92 predictions; range: 25.2)															
MR	18.74	19.36	21.7	22.42	20.12	21.96	24.01	RF	19.11	19.35	19.67	19.74	19.11	19.77	19.43
SVM	18.97	21.24	21.54	19.20	18.60	18.93	18.96	NN	19.76	27.84	23.57	22.77	24.14	21.08	23.79
EA	19.11	19.82	20.41	20.21	19.35	20.35	20.66								
Panel C: Prediction of AAll positive values (93 predictions; range: 37.89)															
MR	12.52	19.26	19.43	20.18	13.76	14.52	15.83	RF	13.52	18.50	19.61	18.58	13.03	13.74	13.52
SVM	14.54	18.31	19.49	18.29	14.19	13.55	14.36	NN	13.41	18.63	21.80	21.39	14.50	17.32	19.07
EA	12.96	18.65	19.09	18.62	13.22	13.86	14.19								
Panel D: Prediction of AAll negative values (93 predictions; range: 25.65)															
MR	16.42	18.26	17.17	18.77	17.62	16.56	18.07	RF	16.64	17.86	16.23	15.90	16.74	15.57	15.44*
SVM	16.77	18.30	16.33	15.88	17.18	17.09	16.37	NN	18.05	22.61	18.34	18.05	17.59	19.81	22.63
EA	16.51	18.04	17.04	16.74	16.63	15.97	16.20								
Panel A: Prediction of II values calculated by BI formula (92 predictions; range: 55.8)															
MR	5.85	12.56	12.12	12.43	6.35	6.49	7.29	RF	8.60	12.14	11.18	11.06	8.37	7.78	8.13
SVM	6.33	14.64	13.10	13.57	6.80	6.79	7.50	NN	8.43	13.18	11.83	12.50	10.95	7.82	9.19
EA	6.22	12.49	11.38	11.02	7.17	6.58	7.39								
Panel B: Prediction of II values calculated by VA formula (91 predictions; range: 19.4)															
MR	16.20	17.48	17.69	19.00	18.12	17.43	19.17	RF	16.7	17.76	16.44	16.78	17.06	16.23	16.59
SVM	16.20	20.27	17.08	16.82	17.63	17.23	16.58	NN	18.16	22.27	15.91	16.95	20.61	17.83	18.97
EA	16.06	17.76	16.2	16.86	16.94	16.43	17.64								
Panel C: Prediction of II positive values (92 predictions; range: 37.9)															
MR	8.31	14.61	12.69	13.31	8.76	8.48	9.11	RF	11.12	14.31	12.74	12.49	10.78	10.24	10.43
SVM	8.76	19.64	13.97	14.04	11.10	8.56	10.39	NN	8.73	16.89	14.11	13.89	12.73	9.67	11.84
EA	8.84	14.71	12.76	12.88	9.17	8.36	8.97								
Panel D: Prediction of II negative values (92 predictions; range: 21.8)															
MR	4.70	11.97	11.93	12.12	5.05	5.53	5.80	RF	6.17	11.31	11.18	10.86	6.23	6.12	6.47
SVM	6.38	13.86	12.26	11.59	9.25	7.00	7.70	NN	6.07	11.42	10.85	11.58	7.39	7.26	8.08
EA	4.87	11.99	10.94	11.66	5.23	5.32	6.04								

For each survey sentiment indicator, the baseline model is underlined and NMAE values lower than baseline are in **bold** (* – p-value < 10%, ** – p-value < 5% and *** – p-value < 1%, NMAE values in %)

The utilization of Twitter indicators produced by AA approach was important for the prediction of VA and negative AAll values. The combined utilization of weekly Twitter indicators and previous AAll values (MSv5) produced lower NMAE results than baseline for VA values. However, the relevance of Twitter indicators is higher for AAll negative values because there are nine models producing lower NMAE results than baseline, four of which use only Twitter indicators. These nine models apply daily Twitter indicators. Furthermore, SVM MSv7 is significantly more accurate than baseline according to the pairwise DMST test. In the prediction of II indicators, there is only one model more accurate than AR(5) model. The utilization of daily Twitter indicators (i.e., MSv3) produces lower NMAE values than baseline for the forecasting of II values calculated by VA formula. For demonstration purposes, the prediction of

negative values of AAI by the SVM MSv4 model (using only Twitter indicators) is presented in the left of Figure 21.

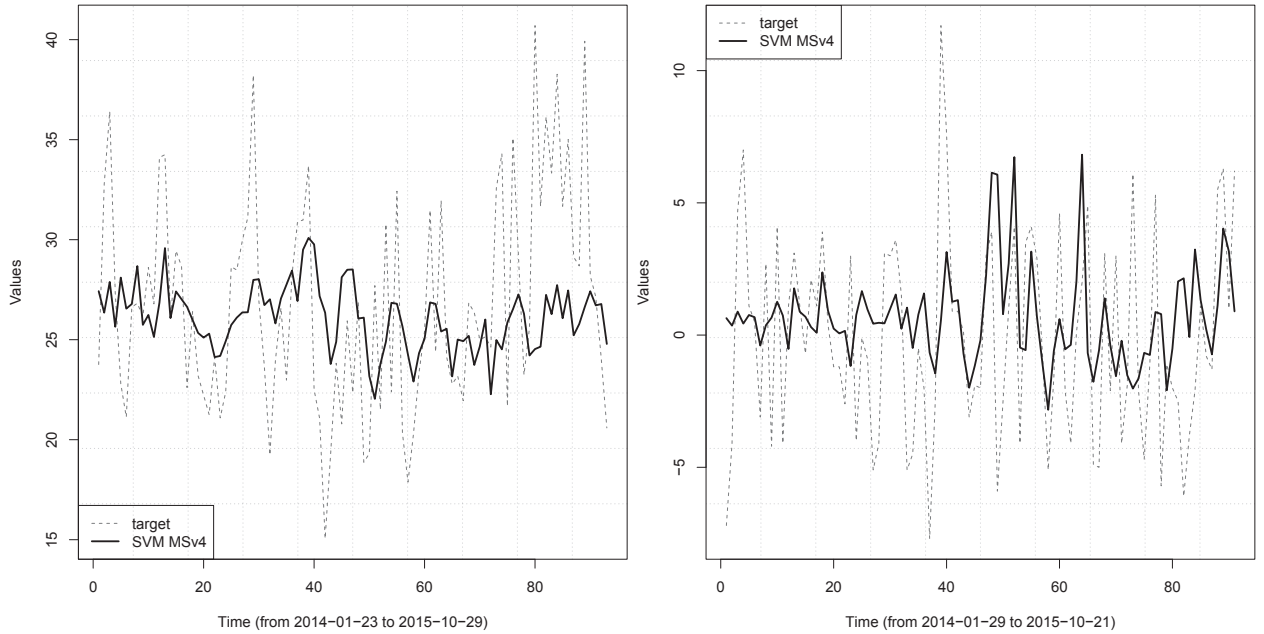


Figure 21.: Predicted results for negative values of AAI (left) and values of II calculated by VA formula using KF indicators (right)

Table 20 shows the results of the prediction of survey indicators using Twitter indicators created by MA approach (average of the daily indicators). The most accurate models using Twitter indicators computed by MA approach produce slightly lower NMAE values than models applying Twitter indicators calculated by AA approach for the forecasting of AAI BI and VA values. However, there are no models significantly more accurate than baseline for the prediction of AAI negative values. Moreover, models utilizing these Twitter indicators do not outperform AR(5) models for the forecasting of any II value.

Table 21 presents the forecasting of survey values using KF indicators. The usage of KF indicators produced worse results than Twitter indicators in the forecasting of AAI indicators. Yet, KF indicators were more informative for the prediction of II VA values. While there is only one model applying Twitter indicators outperforming the baseline model for II values, there are twelve models using KF values more accurate

Chapter 4. The Impact of Microblogging Data for Stock Market Prediction

Table 20.: Prediction of weekly AAll and II values; using Twitter sentiment indicators calculated by MA approach

Mtd	MSv1	MSv2	MSv3	MSv4	MSv5	MSv6	MSv7	Mtd	MSv1	MSv2	MSv3	MSv4	MSv5	MSv6	MSv7
Panel A: Prediction of AAll values calculated by BI formula (93 predictions; range: 58.21)															
MR	<u>14.35</u>	18.81	17.72	19.86	15.93	16.19	17.05	RF	14.61	17.80	17.96	17.71	14.26	15.14	14.85
SVM	14.30	16.44	16.84	17.35	14.40	16.42	15.36	NN	16.11	17.35	17.92	22.71	16.23	17.87	17.58
EA	14.79	17.41	17.12	18.27	14.85	16.53	15.22								
Panel B: Prediction of AAll values calculated by VA formula (92 predictions; range: 25.2)															
MR	<u>18.74</u>	19.65	21.70	23.09	19.89	21.96	23.12	RF	19.11	19.30	19.70	19.58	19.33	19.87	19.34
SVM	18.97	18.92	20.96	19.49	19.11	18.44	18.97	NN	19.76	20.21	23.10	21.53	25.50	22.37	22.15
EA	19.11	19.24	20.11	20.84	19.8	20.35	20.58								
Panel C: Prediction of AAll positive values (93 predictions; range: 37.89)															
MR	<u>12.52</u>	20.65	19.43	20.98	14.15	14.52	16.19	RF	13.52	19.41	19.59	19.67	13.32	13.71	14.01
SVM	14.54	19.38	18.64	19.39	14.07	14.36	14.60	NN	13.41	21.28	20.71	21.11	16.37	15.34	18.72
EA	12.96	19.82	19.25	19.9	14	13.99	14.69								
Panel D: Prediction of AAll negative values (93 predictions; range: 25.65)															
MR	16.42	18.51	17.17	18.86	18.11	16.56	18.06	RF	16.64	16.97	16.35	16.06	16.56	15.45	15.60
SVM	<u>16.77</u>	17.19	16.33	16.30	17.36	17.07	16.19	NN	18.05	18.26	17.80	18.79	17.70	18.70	17.73
EA	16.51	17.33	16.25	16.33	17.08	16.21	15.53								
Panel A: Prediction of II values calculated by BI formula (92 predictions; range: 55.8)															
MR	<u>5.85</u>	12.82	12.12	12.65	6.26	6.49	7.31	RF	8.60	11.87	11.16	10.74	8.14	7.66	8.08
SVM	6.33	14.80	13.32	12.27	10.37	7.11	7.82	NN	8.43	14.61	12.51	12.84	7.37	7.19	10.07
EA	6.22	12.44	11.15	10.98	6.99	6.57	7.65								
Panel B: Prediction of II values calculated by VA formula (91 predictions; range: 19.4)															
MR	16.20	18.14	17.69	18.52	17.96	17.43	17.97	RF	16.70	17.73	16.38	16.92	17.00	16.39	16.58
SVM	16.20	17.78	17.37	19.29	16.85	17.23	18.70	NN	18.16	18.90	16.27	19.17	19.43	20.04	19.48
EA	16.06	18.54	16.33	18.11	17.51	16.78	17.28								
Panel C: Prediction of II positive values (92 predictions; range: 37.9)															
MR	8.31	14.44	12.69	13.45	8.84	8.48	9.25	RF	11.12	14.41	12.80	13.14	10.83	10.12	10.77
SVM	8.76	20.15	14.54	14.24	11.44	8.68	10.51	NN	8.73	15.21	14.42	14.83	9.78	9.94	11.73
EA	8.84	14.55	12.69	12.97	9.64	8.55	9.39								
Panel D: Prediction of II negative values (92 predictions; range: 21.8)															
MR	4.70	11.98	11.93	12.05	4.88	5.53	5.82	RF	6.17	11.11	11.06	10.74	6.12	6.07	6.43
SVM	6.38	12.91	12.25	11.47	9.13	5.47	7.79	NN	6.07	11.43	10.90	11.51	6.39	6.97	8.38
EA	4.87	11.87	11.27	11.51	5.5	5.4	6.22								

For each survey sentiment indicator, the baseline model is underlined and NMAE values lower than baseline are in **bold** (* – p-value < 10%, ** – p-value < 5% and *** – p-value < 1%, NMAE values in %)

than the best AR(5) model for II values calculated by VA approach. These II forecasts are presented in the right of Figure 21.

Overall, Twitter sentiment indicators proved to be useful for the prediction of negative values of AAll but less important for the forecasting of II values. Nevertheless, KF indicators are informative for the prediction of II computed by VA formula. Contrary to the forecasting of stock market variables, SVM is not the dominant regression model and the most accurate AAll and II prediction models are obtained by distinct learning models (e.g., RF is the best model for negative AAll predictions; NN obtains the lowest NMAE values for the prediction of II calculated by VA formula).

Table 21.: Prediction of weekly AAI and II values using KF sentiment indicators

Mtd	MSv1	MSv2	MSv3	MSv4	MSv5	MSv6	MSv7	Mtd	MSv1	MSv2	MSv3	MSv4	MSv5	MSv6	MSv7
Panel A: Prediction of AAI values calculated by BI formula (93 predictions; range: 58.21)															
MR	<u>14.35</u>	18.96	17.77	19.1	15.22	15.69	17.73	RF	14.61	17.34	18.73	17.68	14.36	15.84	15.27
SVM	<u>14.3</u>	18.12	16.36	17.69	14.58	14.88	15.32	NN	16.11	20.77	17.11	19.14	17.11	17.38	19.64
EA	14.79	18.29	16.86	16.87	14.66	14.83	15.58								
Panel B: Prediction of AAI values calculated by VA formula (92 predictions; range: 25.2)															
MR	18.74	20.48	20.82	23.76	20.55	20.86	23.55	RF	19.11	19.67	20.75	20.59	18.95	20.2	19.87
SVM	18.97	20.93	21.79	26.9	19.27	21.25	20.07	NN	19.76	21.2	21.12	24.34	23.24	19.61	21.89
EA	19.11	20.13	19.88	21.07	19.44	19.34	20.34								
Panel C: Prediction of AAI positive values (93 predictions; range: 37.89)															
MR	<u>12.52</u>	20.47	19.94	21.87	13.6	14.16	15.94	RF	13.52	19.41	20.35	20.14	12.8	14.06	13.83
SVM	14.54	18.79	21	20.15	13.42	15.43	14.38	NN	13.41	23.29	20.46	25.15	15.23	15.86	15
EA	12.96	19.83	19.44	19.85	13.24	13.68	14.39								
Panel D: Prediction of AAI negative values (93 predictions; range: 25.65)															
MR	<u>16.42</u>	18.63	16.76	17.67	18.11	16.39	18.13	RF	16.64	17.34	17.04	16.45	16.55	16.48	16.33
SVM	<u>16.77</u>	18.88	16.13	16.56	17.85	18.41	16.59	NN	18.05	18.94	16.86	19.99	20.05	20.21	20.17
EA	16.51	18.11	15.81	16.09	17.75	15.95	15.81								
Panel A: Prediction of II values calculated by BI formula (92 predictions; range: 55.8)															
MR	<u>5.85</u>	11.74	11.83	11.44	6.12	6.44	6.81	RF	8.6	10.95	10.55	10.13	8.06	7.71	7.88
SVM	6.33	14.05	11.56	10.88	8.61	6.55	7.02	NN	8.43	13.09	11.62	12.41	6.81	7.28	7.93
EA	6.22	11.42	10.56	10.51	6.96	6.79	6.53								
Panel B: Prediction of II values calculated by VA formula (91 predictions; range: 19.4)															
MR	16.2	15.70	16.72	17.19	15.99	17.44	17.99	RF	16.7	15.67	16.77	16.18	15.75	16.55	16.27
SVM	16.2	16.36	16.57	15.43	16.58	15.74	16.00	NN	18.16	17.36	19.4	16.46	16.73	17.4	17.87
EA	<u>16.06</u>	15.33	16.53	15.95	16.00	15.89	15.56								
Panel C: Prediction of II positive values (92 predictions; range: 37.9)															
MR	8.31	14.54	13.39	13.41	8.65	8.71	9.36	RF	11.12	13.85	12.99	12.45	11.2	10.34	10.7
SVM	8.76	16.19	13.39	12.78	9.77	8.96	9.02	NN	8.73	16.13	14.6	13.81	10.08	9.85	11.21
EA	8.84	14.4	12.85	12.37	9	8.35	9.17								
Panel D: Prediction of II negative values (92 predictions; range: 21.8)															
MR	4.7	10.81	10.95	10.58	5.09	5.21	5.45	RF	6.17	10.16	9.59	9.41	6.12	6.03	6.18
SVM	<u>6.38</u>	11.33	10.91	10.47	5.74	7.31	7.99	NN	6.07	13.07	9.22	10.6	5.9	6.12	6.62
EA	4.87	10.09	9.91	10.07	5.31	5.46	5.96								

For each survey sentiment indicator, the baseline model is underlined and NMAE values lower than baseline are in **bold** (* – p-value < 10%, ** – p-value < 5% and *** – p-value < 1%, NMAE values in %)

4.8.5 Summary of the Predictive Results

Table 22 summarizes the results obtained in the prediction of stock market variables and survey sentiment indices by indicating the number of models that are significant in the pairwise DMST test. For each predicted variable, it shows the number of these models having TWT (*TWT*), KF (*KF*), posting volume (*Nt*) information and the respective methods (*Method*). To facilitate the analysis, the non-significant predictive models are not shown in the table.

Microblogging sentiment (*TWT*) indicators and posting volume information (*Nt*) were especially informative for the forecasting of SP500, portfolios of lower market capitalization (i.e., Lo20 and Lo30) and some industries (e.g., High Technology, Energy and Telecommunications). There were diverse models that were significant in the

Table 22.: Summary of prediction results (number of models with a p-value < 10% in DMST test for each predicted variable with text based indicators)

Pred. Variables	Indicators				Pred. Variables	Indicators			
	TWT	KF	Nt	Method ^a		TWT	KF	Nt	Method ^a
Returns of indices									
DJIA	1	1	0	SVM(2)	MOM	0	1	0	SVM(1)
RMRF	0	1	0	SVM(1)	SMB	1	0	1	SVM(1)
SP500	2	1	4	SVM(5)					
Returns of portfolios formed on size									
Lo30	4	0	4	SVM(4)	Lo20	8	2	8	SVM(11),EA(1)
Returns of portfolios formed on industries (general sentiment)									
Enrgy	6	4	6	SVM(11),NN(1)	Hitec	3	3	3	SVM(6)
Other	2	1	3	SVM(4)	Shops	1	0	1	SVM(1)
Telcm	1	0	0	SVM(1)					
Returns of portfolios formed on industries (sectorial sentiment)									
Enrgy	4	0	2	SVM(4)	Hitec	7	0	4	SVM(6),EA(2)
Shops	3	0	4	SVM(4)	Telcm	3	0	3	SVM(3)
Trading volume									
DJIA	2	0	2	SVM(2)					
Volatility									
DJIA	0	1	0	MR(1)					
Survey Sentiment Indices									
AAll(neg)	1	0	-	RF(1)					

a - $M_L(n)$, where M_L denotes the ML method (MR, NN, SVM, RF and EA) and n is the number of models using such method.

DMST test for the prediction of these variables. These results support previous findings about the informative content of social media sentiment and posting volume for the forecasting of returns (e.g., Sabherwal et al., 2011; Sprenger et al., 2014; Bollen et al., 2011) and the higher impact of sentiment on smaller stocks (e.g., Baker and Wurgler, 2006; Baker et al., 2012). We highlight that we predicted short term returns of portfolios formed on size, while most studies forecasted medium to long term returns and used other sentiment sources (e.g., financial data, surveys).

The sectorial indicators seem to benefit the prediction of the returns of some industries. Several models applying these indicators for the prediction of returns from Shops, Telecommunications and High Technology portfolio sectors were significantly better, considering the DMST test, when compared against using general sentiment indicators.

The utilization of KF indicators was particularly useful for the prediction of returns of portfolios of small size stocks and the High Technology and Energy sectors. When forecasting returns, there are several models using KF values that provide statistically better results than the baseline model (e.g., Lo20 (2 models), Enrgy (4 models), HiTec (3 models)) according to the DMST test.

Microblogging data was less relevant for the prediction of trading volume and volatility. For these variables there is only evidence of better forecasting ability for the DJIA index.

Regarding the forecasting of survey sentiment values, microblogging sentiment was particularly informative for the prediction of the negative values of AAI. A RF model using Twitter sentiment calculated by AA approach was significant in the DMST test for the forecasting of these survey values. Additionally, there are diverse models using Twitter sentiment and KF indicators that produced lower NMAE values than the most accurate AR(5) model for the prediction of the negative values of AAI and II values computed by VA formula. Therefore, we consider that Twitter and KF indicators can be informative to predict AAI and II values and produce an adequate anticipation or a satisfactory alternative of these survey sentiment indices with some advantages (e.g., cheaper, faster, higher periodicities, targeted to specific stocks).

4.9 CONCLUSIONS

In this work, we proposed a robust methodology that allows to assess the usefulness of microblogging data to the prediction of stock market variables. The methodology (shown in Figure 18) assumes the usage of sentiment and attention indicators extracted from microblogs and survey indices, diverse forms of a daily aggregation of these indicators, usage of KF to merge microblog and survey data sources, a realistic rolling windows evaluation, several ML methods and the DMST test to check if the sentiment and attention based predictions are useful when compared with an autoregressive baseline. In particular, a very recent and large Twitter data set was collected and adopted, with around 31 million tweets from December 2012 to October 2015, related with around 3,800 stocks traded in US markets. The Twitter sentiment indicators were extracted by considering a recent lexicon specifically adjusted to financial microblogging data (Oliveira et al. (2016) and Chapter 3). To the best of our knowledge, this is the first

study that uses sentiment indicators created from specialized financial microblogging lexicons. Moreover, it uses a much larger data period than the majority of studies using Twitter data to predict stock market behavior. Furthermore, we created a novel daily sentiment indicator based on the KF procedure that allows the combination of a daily Twitter indicator with weekly AAI and II values and monthly UMCS and Sentix values. The predictive content of this KF indicator was compared with the content of Twitter indicator. We also explored different sentiment aggregation formulas, such as BullR, VA and VA. We predicted daily returns, trading volume and volatility of diverse indices, such as SP500, RSL, DJIA and NDQ, and portfolios (e.g., formed on size and industries). A fixed-sized rolling window of 300 training days was applied to predict the next day, allowing to perform a large number of model trainings and predictions (ranging from 392 to 439). Also, we explored five different ML methods: MR, NN, SVM, RF and EA. To analyze the predictive value of microblogging features, we considered the NMAE and applied the DMST test between the most accurate AR(5) model (baseline) and similar models using microblogging extracted variables. In our opinion, this is a more robust methodology when compared with most state of the art works.

Additionally, some state of the art studies have analyzed the influence of sentiment on portfolios formed on diverse characteristics (e.g., market capitalization, book-to-market ratio). Many of these works refer that the effect of sentiment is more evident on returns of some portfolios having extreme values (e.g., small market capitalization (Baker and Wurgler, 2006; Baker et al., 2012)). However, most of these studies apply low frequencies (e.g., monthly, annual) and do not use sentiment indicators extracted from social media. Therefore, we analyzed in this study the predictive content of daily sentiment indicators extracted from Twitter on returns of portfolios formed on size and industries.

Considering that AAI and II are widely used by academics (Solt and Statman, 1988; Fisher and Statman, 2000; Verma and Soydemir, 2009) and practitioners, we also predicted these sentiment measures using Twitter and KF sentiment indicators. To the best of our knowledge, this is the first study that addresses such forecasting.

We found that microblogging sentiment and attention indicators were particularly useful for the prediction of returns of SP500 index, portfolios of lower market capitalization and some sectors such as High Technology, Energy and Telecommunications. In these situations, there are models obtaining p-value less than 5% in the pairwise DMST test with the baseline model. The application of microblogging features were less convincing for the forecasting of trading volume and volatility. These results add some evidence about the predictive content of social media sentiment and posting volume for returns (e.g., Sabherwal et al., 2011; Sprenger et al., 2014; Bollen et al., 2011). Additionally, microblogging sentiment indicators have various advantages when compared to traditional sentiment measures (e.g., surveys). For instance, their creation is faster and cheaper, allows greater frequencies (e.g., daily) and may be targeted to a more limited group of stocks (e.g., individual stocks or indices). The obtained results also corroborate previous findings that sentiment has more impact on smaller stocks (e.g., Baker and Wurgler, 2006; Baker et al., 2012). We note that daily microblogging features were used in this study while the majority of previous studies apply monthly sentiment indicators extracted from economic variables or surveys.

The utilization of sentiment indicators produced by KF were informative for the prediction of returns of some portfolios and indices. There are models using KF indicators having p-value less than 5% in the pairwise test for SP500, Lo20, HiTec and Enrgy. Also, the application of KF indicators decreased NMAE values for diverse indices and portfolios. However, KF indicators were less effective for the forecasting of trading volume and volatility.

Twitter sentiment values were specially informative for negative values of AAIL. In this case, there were several models applying Twitter indicators producing lower NMAE results than the most accurate AR(5) model. One of these models is significantly more accurate than baseline according to the pairwise DMST test, and uses exclusively Twitter indicators. KF indicators were particularly important for the prediction of II values calculated by VA formula. In this situation, there were twelve models more accurate than the baseline model. These results show that Twitter and

KF indicators can be valuable to forecast AAI and II values. Therefore, they may permit a satisfactory anticipation of these sentiment indicators or an acceptable alternative whenever they are unavailable.

There are several finance studies that show that sentiment can explain medium to long term returns (e.g., Baker and Wurgler, 2006; Baker et al., 2012), but there is scarce evidence that sentiment can forecast short term returns. Our results are of interest for academics and practitioners alike. This research contributes to the literature by providing evidence that sentiment extracted from microblogging has short term predictive power for some series of financial returns. These results open future research avenues, for instance to test if firms are more prone to being affected by sentiment in times of greater ambiguity and risk (for instance around Initial Public Offerings). Moreover, the results obtained when forecasting sentiment surveys show that sentiment can be used to predict their future values. And they also justify trying to explore the applied methodology in other business areas, namely in marketing, as sentiment can be used as a proxy for consumer satisfaction.

Thus, the proposed methodology could be adopted by financial decision support systems to assist investors in their decisions by providing instant access to social media analytics, such as customized sentiment indicators or predictions. In the future, we also intend to identify influential microblog users and assess their contribution to the forecasting of specific stocks. For instance, social influence analysis has been applied for advertisement purposes (Gundecha and Liu, 2012) but has been scarcely explored for the stock markets. However, the identification of influent social media users may allow the creation of sentiment indicators of informed users and anticipate the overall sentiment of investors. Additionally, most studies apply only one source of Web data so the complementarity value of different data sources remains unclear. These sources have distinct characteristics that can be complementary and enable better predictions. For instance, blogs have more complete opinionated content, microblogging contents have greater objectivity, interactivity and posting frequencies and Google searches rep-

resent a superior number of users. The dynamic combination of diverse Web data sources may result in more informative financial indicators.

Part III

CONCLUSIONS

CONCLUSIONS

5.1 OVERVIEW

Stock market agents have been increasingly using social media services to share important information related to their activities. For instance, microblogs are a recent popular platform where many investors frequently publish valuable information such as their opinions, their stock market operations or new information about stocks. Hence, microblogging data may be an informative data source for stock market prediction. For example, the vast opinionated content of these data can allow a fast creation of a wide range of representative investor sentiment indicators. Furthermore, these indicators present several advantages when compared with traditional sentiment indices (e.g., surveys). The production of microblogging sentiment indicators is more rapid, economical and flexible. Financial decision support systems may provide real-time customized (e.g., personalized periodicity and stocks) sentiment and attention values extracted from microblogging data. Thus, microblogging data may allow a more efficient creation of predictive models for stock market behavior.

However, it is essential to evaluate the predictive power of microblogging data for stock market behavior in order to assess their real utility. Diverse studies have analyzed the impact of sentiment and attention on stock market prediction (e.g., Antweiler and Frank, 2004; Das and Chen, 2007; Bollen et al., 2011; Sprenger et al., 2014; Oh and

Sheng, 2011). Various proxies for investor sentiment have been used such as financial variables, surveys and textual contents (e.g., newspapers, message boards, microblogs).

Diverse studies have found that sentiment is useful to make trading decisions (Schumaker et al., 2012; Chen and Lazer, 2013) or forecast stock market variables, such as stock prices (Tetlock, 2007; Bollen et al., 2011), price movements (Groß-Klußmann and Hautsch, 2011; Oh and Sheng, 2011), returns (Sprenger et al., 2014; Bollen et al., 2011; Garcia, 2013), volatility (Antweiler and Frank, 2004; Sabherwal et al., 2011) and trading volume (Tetlock, 2007; Antweiler and Frank, 2004). Moreover, posting volume on social media services (e.g., microblogs and message boards) has also been successfully applied in the forecasting of returns (e.g., Wysocki, 1998; Antweiler and Frank, 2004), trading volume (e.g., Sprenger et al., 2014; Wysocki, 1998) and volatility (e.g., Antweiler and Frank, 2004; Das and Chen, 2007).

Nonetheless, there are diverse research challenges in this topic. For instance, the research findings are controversial because there are several studies that find little evidence of the predictive value of sentiment or attention for stock market variables (Tumarkin and Whitelaw, 2001; Antweiler and Frank, 2004; Das et al., 2005; Timmermann, 2008). The majority of the studies present limitations in their evaluation procedure. For example, most studies do not perform out of sample evaluation and rely exclusively on the linear models MR or VAR to forecast stock market variables. The utilization of more flexible learning models (e.g., SVM, NN) is limited and the comparison of these models is very scarce. The size of social media data sets has been very low (e.g., less than two years), the applied test sets are very short and the utilization of statistical tests to evaluate the predictive accuracy is very scant.

The evaluation of the predictive content of sentiment on portfolios (e.g., volatility, book to market, size) is common on surveys or financial data studies but it is absent for social media and for higher periodicities than weekly. The lexicons applied in the creation of sentiment indicators are not properly adapted to microblogging stock market contents. Moreover, the various investor sentiment indicators used in this topic may contain some noise because they have different characteristics (e.g., sources, fre-

quencies) and commonly present distinct values. The aggregation of these sentiment indicators may create a more representative investor sentiment value. To the best of our knowledge, there are no studies combining social media indicators with other sentiment sources.

The forecasting of survey sentiment indices can be valuable for investors because it may permit an adequate anticipation or be a inexpensive alternative measure. Nonetheless, there is no previous research focused on forecasting these indices.

The main objective of this thesis was to perform a rigorous evaluation of the utility of microblogging data for the prediction of stock market behavior. To satisfy this research objective, this work addressed the identified challenges in this topic (last row of Table 2). Our predictive models used microblog sentiment and attention indicators extracted from a larger Twitter data set than most social media studies use. The lack of lexicons properly adapted to microblogging stock market contents was solved by the proposal of a procedure to automatically create a specialized lexicon from labeled StockTwits data. This produced lexicon permitted a fast and inexpensive unsupervised creation of reliable investor sentiment indicators from a Twitter data set. Moreover, we experimented the extraction of an unique daily sentiment indicator from a daily microblogging indicator, two weekly survey indices (AAII, II) and two monthly survey measures (UMCS, Sentix) by applying a KF procedure. Five distinct ML models were tested in the forecasting of stock market variables associated to indices and portfolios and in the prediction of two popular weekly sentiment survey indices (AAII and II). The predictive value of microblogging data was assessed by the pairwise DMST test.

In summary, the main contributions of this PhD dissertation are:

- execute a robust evaluation of the utility of microblogging data for the forecasting of stock market variables by applying a more solid methodology than used in this topic;
- perform a large number of experiments to analyze the potential different predictive power of sentiment for distinct stock market variables (e.g., returns, volatility

and trading volume) and types (e.g., indices, portfolios). For instance, the forecasting of portfolios daily returns using social media data is very limited;

- present an original procedure to automatically produce a specialized microblogging stock market lexicon. There were no lexicons properly adjusted to microblogging messages about stock market. The created lexicon should allow a fast and light unsupervised creation of more robust investor sentiment indicators. The proposed approach applies three adapted statistical measures (e.g., PMI) and two novel complementary statistics on labeled StockTwits messages to compute a stock market sentiment score. Additionally, this procedure calculates distinct sentiment values for affirmative and negated contexts in order to address negation more effectively;
- generate new robust microblogging sentiment indicators based on the newly produced microblogging stock market lexicon and a larger Twitter data set than most studies using social media data. Furthermore, we tested the creation of sentiment indicators for diverse targets (e.g., aggregated market, industries) and aggregated by different formulas (e.g., BullR, VA, AG);
- extract an unique daily sentiment indicator from measures of different frequencies (i.e., daily, weekly, monthly) and sources (Twitter, AAIL, II, UMCS, Sentix). This indicator created by a KF procedure may constitute a less noisy measure of investor sentiment;
- forecast existing survey sentiment indices (AAII and II) by applying microblogging data, which may permit an acceptable anticipation of those values or a satisfactory alternative whenever they are unavailable.

5.2 DISCUSSION

The experimental work of this thesis started with the creation of a specialized stock market lexicon properly adapted to microblogging stock market messages. There are very few financial lexicons and the existing ones were created using distinct data sources. Therefore, we considered that it was important to produce a specialized lexicon that could allow a fast and inexpensive unsupervised creation of accurate investor sentiment indicators for the following research activities.

The proposed procedure permitted the generation of lexicons that substantially outperform six reference lexicons. The SA results obtained by the created lexicons on StockTwits data were significantly higher than the SA results produced by the baseline lexicons. Moreover, the application of the novel complementary metrics proved to be relevant. Lexicons using any of these metrics significantly improved SA compared to their counterparts that do not apply these measures. Additionally, the usage of affirmative and negated context sentiment values was useful. Lexicons using these dual scores enhanced or maintained almost all evaluation results compared to their counterparts. Furthermore, Twitter based sentiment indicators produced by a microblogging stock market lexicon (SML) showed a significant moderate Pearson's correlation with popular AAI and II survey indices.

A lexicon created by the proposed procedure (SML) was applied in the analysis of the impact of microblogging data for stock market prediction. SA results obtained by this lexicon (shown in Chapter 3) convinced us that its utilization would allow a fast and accessible creation of reliable investor sentiment indicators from a very large unlabeled Twitter data set (approximately 31 million messages).

The experiments revealed that microblogging sentiment and attention indicators were particularly informative for the forecasting of returns of SP500 index, portfolios of lower dimension and some industries such as Energy, High Technology and Telecommunications. The utilization of microblogging features was less relevant for the pre-

diction of volatility and trading volume. These results add some evidence about the predictive value of social media sentiment and attention for returns (e.g., Sabherwal et al., 2011; Sprenger et al., 2014; Bollen et al., 2011).

Furthermore, the obtained results support previous findings that sentiment is more influential on smaller stocks (e.g., Baker and Wurgler, 2006; Baker et al., 2012). We highlight that daily microblogging features were applied in this study while most studies use monthly sentiment indicators extracted from financial variables or surveys (e.g., Baker and Wurgler, 2006; Baker et al., 2012). Our results are important for academics and practitioners alike. This project contributes to the literature by providing evidence that microblogging sentiment has short term predictive value for some series of financial returns.

The application of a unique daily sentiment indicator aggregating microblogging and survey data was informative for the forecasting of returns of some indices (e.g., SP500) and portfolios (e.g., Lo20, HiTec, Enrgy) but it was less important for the prediction of volatility and trading volume.

Twitter sentiment was particularly useful for the forecasting of negative values of AAI while KF indicators were especially informative for the prediction of II values computed by VA formula. Hence, these prediction results demonstrate that microblogging data can be valuable to predict survey sentiment indices. They may allow an adequate anticipation of these values or an appropriate alternative whenever they are unavailable.

The methodology proposed in this thesis could be applied by financial decision support systems to assist users in their stock market decisions by providing instant access to social media analytics, such as personalized sentiment indicators or forecasts. Microblogging sentiment indicators have various advantages when compared to traditional sentiment measures (e.g., surveys). For instance, the extraction of sentiment is more rapid, inexpensive and flexible (e.g., higher frequencies, selected group of stocks).

5.3 FUTURE WORK

The results obtained by this project suggest future research avenues. For instance, the different short term predictive power for diverse series of financial returns may recommend further research about the impact of sentiment on firms in times of higher ambiguity and risk (e.g., around Initial Public Offerings).

The results obtained in the prediction of survey indicators show that microblogging data can have informative content for survey prediction. Therefore, it could be important to extend research about the forecasting of these values. For instance, the application of other data sources or the selection of specific user communities may improve this procedure.

Social media analysis has been applied in diverse areas (e.g., advertisement) to assess users influence. For example, the identification of influential users may permit the maximization of social media advertisement campaigns. The assessment of influence in stock market platforms has been scarcely performed and it may improve the creation of investor sentiment indicators and enhance stock market prediction. An effective identification of these influential investors may allow the anticipation of sentiment trends and the extraction of more representative sentiment indicators for specific stocks.

Furthermore, very few studies apply more than one social media data source to predict stock market behavior. Therefore, the evaluation of the complementarity value of different data sources is very incomplete. We consider that these sources have distinct characteristics that may be complementary. For example, microblogging contents are highly objective, interactive and frequent, blog posts are longer and may be more complete and Google searches represent a larger number of users. The dynamic combination of these data sources may permit richer financial indicators.

The proposed approach to automatically create a stock market lexicon improved SA on microblogging stock market messages. However, this procedure may be enhanced

Chapter 5. Conclusions

by exploring other approaches. For instance, collective intelligence can be applied to more easily label unclassified text and identify relevant stock market terms. So, the proposed procedure could be applied to other stock market message sources (e.g., Twitter) and the created lexicons could be more accurate and extensive. Moreover, the utilization of active learning algorithms may reduce the manual labeling effort by automatically selecting a small set of important messages for human classification.

REFERENCES

- Alya Al Nasser, Allan Tucker, and Sergio de Cesare. Quantifying stocktwits semantic terms trading behavior in financial markets: An effective application of decision tree algorithms. *Expert Systems with Applications*, 42(23):9192–9210, 2015.
- Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 579–586. Association for Computational Linguistics, 2005.
- Yali Amit and Donald Geman. Shape quantization and recognition with randomized trees. *Neural computation*, 9(7):1545–1588, 1997.
- Michelle Annett and Grzegorz Kondrak. A comparison of sentiment analysis techniques: Polarizing movie blogs. In *Conference of the Canadian Society for Computational Studies of Intelligence*, pages 25–35. Springer, 2008.
- Werner Antweiler and Murray Z. Frank. Is all that talk just noise? the information content of internet stock message boards. *The Journal of Finance*, 59(3):1259–1294, 2004.
- Nelson Areal and Stephen J Taylor. The realized volatility of ftse-100 futures prices. *Journal of Futures Markets*, 22(7):627–648, 2002.
- J Scott Armstrong. Evaluating forecasting methods. In *Principles of forecasting*, pages 443–472. Springer, 2001.
- J Scott Armstrong and Fred Collopy. Error measures for generalizing about forecasting methods: Empirical comparisons. *International journal of forecasting*, 8(1):69–80, 1992.

References

- Sitaram Asur and Bernardo A Huberman. Predicting the future with social media. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 1, pages 492–499. IEEE, 2010.
- George S Atsalakis and Kimon P Valavanis. Surveying stock market forecasting techniques—part ii: Soft computing methods. *Expert Systems with Applications*, 36(3):5932–5941, 2009.
- Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation LREC10*, volume 0, pages 2200–2204. European Language Resources Association (ELRA), 2010.
- Xue Bai. Predicting consumer sentiments from online text. *Decision Support Systems*, 50(4):732–742, 2011.
- Malcolm Baker and Jeffrey Wurgler. Investor sentiment and the cross-section of stock returns. *The Journal of Finance*, 61(4):1645–1680, 2006.
- Malcolm Baker and Jeffrey Wurgler. Investor Sentiment in the Stock Market. *Journal of Economic Perspectives*, 21(2):129–151, 2007.
- Malcolm Baker, Jeffrey Wurgler, and Yu Yuan. Global, local, and contagious investor sentiment. *Journal of Financial Economics*, 104(2):272–287, May 2012.
- Adam Bermingham and Alan F Smeaton. On using twitter to monitor political sentiment and predict election results. *Sentiment Analysis where AI meets Psychology (SAAIP)*, 2011.
- Gerard Biau. Analysis of a random forests model. *Journal of Machine Learning Research*, 13(Apr):1063–1095, 2012.
- Albert Bifet and Eibe Frank. Sentiment knowledge discovery in twitter streaming data. In *International Conference on Discovery Science*, pages 1–15. Springer, 2010.

- Christopher M Bishop. Pattern recognition. *Machine Learning*, 128, 2006.
- Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.
- Bernhard Boser, Isabelle Guyon, and Vladimir Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992.
- Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- Gregory W. Brown and Michael T. Cliff. Investor sentiment and the near-term stock market. *Journal of Empirical Finance*, 11(1):1–27, 2004.
- Gregory W. Brown and Michael T. Cliff. *Investor Sentiment and Asset Valuation*, 2005.
- Christopher JC Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998.
- Erik Cambria, Tim Benson, Chris Eckl, and Amir Hussain. Sentic prompts: Application of sentic computing to the development of a novel unified framework for measuring health-care quality. *Expert Systems with Applications*, 39(12):10533–10543, 2012.
- Erik Cambria, Bjorn Schuller, Yunqing Xia, and Catherine Havasi. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, 28(2):15–21, 2013.
- Pimwadee Chaovalit and Lina Zhou. Movie review mining: A comparison between supervised and unsupervised classification approaches. In *Proceedings of the 38th annual Hawaii international conference on system sciences*, pages 112c–112c. IEEE, 2005.
- CL Philip Chen and Chun-Yang Zhang. Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Information Sciences*, 275:314–347, 2014.

References

- Min Chen, Shiwen Mao, and Yunhao Liu. Big data: a survey. *Mobile Networks and Applications*, 19(2):171–209, 2014.
- Ray Chen and Marius Lazer. Sentiment analysis of twitter feeds for the prediction of stock market movement. *Stanford Education*, 25, 2013.
- Tsan-Ming Choi, Hing Kai Chan, and Xiaohang Yue. Recent development in big data analytics for business operations and risk management. *IEEE Transactions on Cybernetics*, 47(1):81–92, 2016.
- Yejin Choi and Claire Cardie. Learning with compositional semantics as structural inference for subsentential sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 793–801. Association for Computational Linguistics, 2008.
- Yoonjung Choi, Youngho Kim, and Sung-Hyon Myaeng. Domain-specific sentiment analysis using contextual feature generation. In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, pages 37–44. ACM, 2009.
- San-Lin Chung, Chi-Hsiou Hung, and Chung-Ying Yeh. When does investor sentiment predict stock returns? *Journal of Empirical Finance*, 19(2):217–240, 2012.
- Pilar Corredor, Elena Ferrer, and Rafael Santamaria. Investor sentiment effect in stock markets: Stock characteristics or country-specific factors? *International Review of Economics & Finance*, 27:572–591, 2013.
- Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.
- P. Cortez. Data Mining with Neural Networks and Support Vector Machines using the R/rminer Tool. In P. Perner, editor, *Advances in Data Mining – Applications and Theoretical Aspects, 10th Industrial Conference on Data Mining*, pages 572–583, Berlin, Germany, July 2010. LNAI 6171, Springer.

- Sven F Crone, Michele Hibon, and Konstantinos Nikolopoulos. Advances in forecasting with neural networks? empirical evidence from the nn3 competition on time series prediction. *International Journal of Forecasting*, 27(3):635–660, 2011.
- George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.
- Zhi Da, Joseph Engelberg, and Pengjie Gao. In search of attention. *The Journal of Finance*, 66(5):1461–1499, 2011.
- Nádia FF da Silva, Eduardo R Hruschka, and Estevam R Hruschka. Tweet sentiment analysis with classifier ensembles. *Decision Support Systems*, 66:170–179, 2014.
- Hoa Trang Dang and Karolina Owczarzak. Overview of the tac 2008 opinion question answering and summarization tasks. In *Proc. of the First Text Analysis Conference*, 2008.
- Sanjiv Das and Mike Chen. Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web. *Management Science*, 53(9):1375–1388, 2007.
- Sanjiv Das, Asís Martínez-Jerez, and Peter Tufano. einformation: A clinical study of investor discussion and sentiment. *Financial Management*, 34(3):103–137, 2005.
- Kushal Dave, Steve Lawrence, and David M Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, pages 519–528. ACM, 2003.
- Shangkun Deng, Takashi Mitsubuchi, Kei Shioda, Tatsuro Shimada, and Akito Sakurai. Combining technical analysis with sentiment analysis for stock price prediction. In *Dependable, Autonomic and Secure Computing (DASC), 2011 IEEE Ninth International Conference on*, pages 800–807. IEEE, 2011.
- Peter J Denning. A new social contract for research. *Communications of the ACM*, 40(2): 132–134, 1997.

References

- Ann Devitt and Khurshid Ahmad. Sentiment polarity identification in financial news: A cohesion-based approach. In *ACL*, 2007.
- Francis X Diebold and Roberto S Mariano. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 1995.
- Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000.
- Adnan Duric and Fei Song. Feature selection for sentiment analysis based on content and syntax models. *Decision support systems*, 53(4):704–711, 2012.
- Paul Ekman, Wallace V Friesen, and Phoebe Ellsworth. *Emotion in the human face: Guidelines for research and an integration of findings*. Elsevier, 2013.
- David Enke and Suraphan Thawornwong. The use of data mining and neural networks for forecasting stock market returns. *Expert Systems with applications*, 29(4):927–940, 2005.
- Andrea Esuli and Fabrizio Sebastiani. Determining the semantic orientation of terms through gloss classification. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, pages 617–624. ACM, 2005.
- Eugene F Fama. Market efficiency, long-term returns, and behavioral finance. *Journal of financial economics*, 49(3):283–306, 1998.
- Weiguo Fan and Michael D Gordon. The power of social media analytics. *Communications of the ACM*, 57(6):74–81, 2014.
- Ronen Feldman. Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4):82–89, 2013.
- Elisabetta Fersini, Enza Messina, and Federico Pozzi. Sentiment analysis: Bayesian ensemble learning. *Decision Support Systems*, 68:26–38, 2014.

- Kenneth L. Fisher and Meir Statman. Investor Sentiment and Stock Returns. *Financial Analysts Journal*, 56(2):16–23, 2000.
- Diego Garcia. Sentiment during recessions. *The Journal of Finance*, 68(3):1267–1300, 2013.
- Robin Genuer, Jean-Michel Poggi, and Christine Tuleau. Random forests: some methodological insights. *arXiv preprint arXiv:0811.3619*, 2008.
- Tomer Geva and Jacob Zahavi. Empirical evaluation of an automated intraday stock recommendation system incorporating both market data and textual news. *Decision support systems*, 57:212–223, 2014.
- Stefan Gindl and Johannes Liegl. Evaluation of different sentiment detection methods for polarity classification on web-based reviews. In *Proceedings of the 18th European conference on artificial intelligence*, pages 35–43, 2008.
- Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1:12, 2009.
- Namrata Godbole, Manja Srinivasaiah, and Steven Skiena. Large-scale sentiment analysis for news and blogs. *ICWSM*, 7(21):219–222, 2007.
- Andrew B Goldberg and Xiaojin Zhu. Seeing stars when there aren't many stars: graph-based semi-supervised learning for sentiment categorization. In *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*, pages 45–52. Association for Computational Linguistics, 2006.
- Ken Goldberg, Theresa Roeder, Dhruv Gupta, and Chris Perkins. Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval*, 4(2):133–151, 2001.
- William H Greene. *Econometric analysis (International edition)*. Pearson US Imports & PHIPES, 2000.

References

- Axel Groß-Klußmann and Nikolaus Hautsch. When machines read the news: Using automated text analytics to quantify high frequency news-implied market reactions. *Journal of Empirical Finance*, 18(2):321–340, 2011.
- Pritam Gundecha and Huan Liu. Mining social media: a brief introduction. *Tutorials in Operations Research*, 1(4):1–17, 2012.
- Michael Hagenau, Michael Liebmann, and Dirk Neumann. Automated news reading: Stock price prediction based on financial news using context-capturing features. *Decision Support Systems*, 55(3):685–697, 2013.
- Jiawei Han, Micheline Kamber, and Jian Pei. *Data mining: concepts and techniques*. Elsevier, 2011.
- Jonghyun Han and Hyunju Lee. Characterizing the interests of social media users: Refinement of a topic model for incorporating heterogeneous media. *Information Sciences*, 358:112–128, 2016.
- Lars Kai Hansen and Peter Salamon. Neural network ensembles. *IEEE transactions on pattern analysis and machine intelligence*, 12:993–1001, 1990.
- Trevor Hastie, Robert John Tibshirani, and Jerome H Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, NY, USA, 2nd edition, 2008.
- Trevor Hastie, Robert John Tibshirani, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2013.
- Vasileios Hatzivassiloglou and Kathleen McKeown. Predicting the semantic orientation of adjectives. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 174–181, 1997.

- Vasileios Hatzivassiloglou and Janyce M Wiebe. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of the 18th conference on Computational linguistics-Volume 1*, pages 299–305. Association for Computational Linguistics, 2000.
- Yulan He and Deyu Zhou. Self-training from labeled features for sentiment analysis. *Information Processing & Management*, 47(4):606–616, 2011.
- Bas Heerschop, Frank Goossen, Alexander Hogenboom, Flavius Frasinca, Uzay Kaymak, and Franciska de Jong. Polarity analysis of texts using discourse structure. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1061–1070. ACM, 2011.
- Jördis Hengelbrock, Erik Theissen, and Christian Westheide. Market response to investor sentiment. *Journal of Business Finance & Accounting*, 40(7-8):901–917, 2013.
- Alan Hevner and Samir Chatterjee. *Design science research in information systems*. Springer, 2010.
- Alan Hevner, Salvatore T March, Jinsoo Park, and Sudha Ram. Design science in information systems research. *MIS quarterly*, 28(1):75–105, 2004.
- David Hirshleifer and Siew Hong Teoh. Limited attention, information disclosure, and financial reporting. *Journal of accounting and economics*, 36:337–386, 2003.
- Chienwei Ho and Chi-Hsiou Hung. Investor sentiment as conditioning information in asset pricing. *Journal of Banking & Finance*, 33(5):892–903, May 2009.
- Tin Kam Ho. The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, 20(8):832–844, 1998.
- Paul Hribar and John McInnis. Investor sentiment and analysts’ earnings forecast errors. *Management Science*, 58(2):293–307, 2012.

References

- Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.
- Nan Hu, Indranil Bose, Noi Sian Koh, and Ling Liu. Manipulation of online reviews: An analysis of ratings, readability, and sentiments. *Decision Support Systems*, 52(3): 674–684, 2012.
- Zan Huang, Hsinchun Chen, Chia-Jung Hsu, Wun-Hwa Chen, and Soushan Wu. Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decision support systems*, 37(4):543–558, 2004.
- Clayton J Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International AAI Conference on Weblogs and Social Media*, 2014.
- Rob J Hyndman and Anne B Koehler. Another look at measures of forecast accuracy. *International journal of forecasting*, 22(4):679–688, 2006.
- Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. Target-dependent twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 151–160. Association for Computational Linguistics, 2011.
- Jaap Kamps, Robert Mokken, Maarten Marx, and Maarten De Rijke. Using WordNet to measure semantic orientation of adjectives. In *Proceedings of the 4th International Conference on Language Resources and Evaluation LREC 2004*, volume 4, pages 1115–1118. Citeseer, 2004.
- Hanhoon Kang, Seong Joon Yoo, and Dongil Han. Senti-lexicon and improved naïve bayes algorithms for sentiment analysis of restaurant reviews. *Expert Systems with Applications*, 39(5):6000–6010, 2012.

- Alistair Kennedy and Diana Inkpen. Sentiment classification of movie reviews using contextual valence shifters. *Computational intelligence*, 22(2):110–125, 2006.
- Soo-Min Kim and Eduard Hovy. Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics, 2004.
- Svetlana Kiritchenko, Xiaodan Zhu, and Saif M Mohammad. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 50:723–762, 2014.
- Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Stanford, CA, 1995.
- Efstratios Kontopoulos, Christos Berberidis, Theologos Dergiades, and Nick Bassiliades. Ontology-based sentiment analysis of twitter posts. *Expert systems with applications*, 40(10):4065–4074, 2013.
- Lun-Wei Ku, Yu-Ting Liang, and Hsin-Hsi Chen. Opinion extraction, summarization and tracking in news and blog corpora. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 100107, 2006.
- Taku Kudo and Yuji Matsumoto. A boosting algorithm for classification of semi-structured text. In *EMNLP*, volume 4, pages 301–308, 2004.
- P Ravi Kumar and Vadlamani Ravi. Bankruptcy prediction in banks and firms via statistical and intelligent techniques—a review. *European journal of operational research*, 180(1):1–28, 2007.
- Alexander Kurov. Investor sentiment, trading behavior and informational efficiency in index futures markets. *Financial Review*, 43(1):107–127, 2008.
- Alexander Kurov. Investor sentiment and the stock markets reaction to monetary policy. *Journal of Banking & Finance*, 34(1):139–149, January 2010.

References

- Namhee Kwon, Stuart W Shulman, and Eduard Hovy. Multidimensional text analysis for erulemaking. In *Proceedings of the 2006 international conference on Digital government research*, pages 157–166. Digital Government Society of North America, 2006.
- Vasileios Lampos, Daniel Preotiuc-Pietro, and Trevor Cohn. A user-centric model of voting intention from social media. In *ACL (1)*, pages 993–1003, 2013.
- Doug Laney. 3d data management: Controlling data volume, velocity and variety. *META Group Research Note*, 6:70, 2001.
- Michael Laver, Kenneth Benoit, and John Garry. Extracting policy positions from political texts using words as data. *American Political Science Review*, 97(02):311–331, 2003.
- Charles Lee, Andrei Shleifer, and Richard Thaler. Investor sentiment and the closed-end fund puzzle. *The Journal of Finance*, 46(1):75–109, 1991.
- Wayne Y. Lee, Christine X. Jiang, and Daniel C. Indro. Stock market volatility, excess returns, and the role of investor sentiment. *Journal of Banking and Finance*, 26(12):2277–2299, 2002.
- Michael Lemmon and Evgenia Portniaguina. Consumer confidence and asset prices: Some empirical evidence. *Review of Financial Studies*, 19(4):1499–1529, 2006.
- Kevin Lerman, Sasha Blair-Goldensohn, and Ryan McDonald. Sentiment summarization: evaluating and learning user preferences. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 514–522. Association for Computational Linguistics, 2009.
- Cane WK Leung, Stephen CF Chan, and Fu-lai Chung. Integrating collaborative filtering and sentiment analysis: A rating inference approach. In *Proceedings of the ECAI 2006 workshop on recommender systems*, pages 62–66, 2006.

- David D. Lewis. An evaluation of phrasal and clustered representations on a text categorization task. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '92*, pages 37–50, New York, NY, USA, 1992. ACM.
- Qing Li, TieJun Wang, Ping Li, Ling Liu, Qixu Gong, and Yuanzhu Chen. The effect of news and public mood on stock movements. *Information Sciences*, 278:826–840, 2014.
- Yung-Ming Li and Tsung-Ying Li. Deriving market intelligence from microblogs. *Decision Support Systems*, 55(1):206–217, 2013.
- Chenghua Lin and Yulan He. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 375–384. ACM, 2009.
- Dekang Lin. Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 2*, pages 768–774. Association for Computational Linguistics, 1998.
- Wei-Yang Lin, Ya-Han Hu, and Chih-Fong Tsai. Machine learning in financial crisis prediction: a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4):421–436, 2012.
- Lucian Vlad Lita, Andrew Hazen Schlaikjer, WeiChang Hong, and Eric Nyberg. Qualitative dimensions in question answering: Extending the definitional qa task. In *Proceedings of the National Conference on Artificial Intelligence*, volume 4, pages 1616–1617, 2005.
- Bing Liu and Lei Zhang. A survey of opinion mining and sentiment analysis. In *Mining text data*, pages 415–463. Springer, 2012.

References

- Hugo Liu, Henry Lieberman, and Ted Selker. A model of textual affect sensing using real-world knowledge. In *Proceedings of the 8th international conference on Intelligent user interfaces*, pages 125–132. ACM, 2003.
- J. Bradford De Long, Andrei Shleifer, Lawrence H. Summers, and Robert J. Waldmann. Noise Trader Risk in Financial Markets. *Journal of Political Economy*, 98(4):703, 1990.
- Tim Loughran and Bill McDonald. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance*, 66(1):35–65, 2011.
- Yue Lu, Malu Castellanos, Umeshwar Dayal, and ChengXiang Zhai. Automatic construction of a context-aware sentiment lexicon: an optimization approach. In *Proceedings of the 20th international conference on World wide web*, pages 347–356. ACM, 2011.
- Masoud Makrehchi, Shalin Shah, and Wenhui Liao. Stock prediction using event-based sentiment analysis. In *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on*, volume 1, pages 337–342. IEEE, 2013.
- Huina Mao, Scott Counts, and Johan Bollen. Predicting financial markets: Comparing survey, news, twitter and search engine data. *arXiv preprint arXiv:1112.1051*, 2011.
- Huina Mao, Pengjie Gao, Yongxiang Wang, and Johan Bollen. Automatic construction of financial semantic orientation lexicon from large-scale chinese news corpus. In *7th Financial Risks International Forum*. Institut Louis Bachelier, 2014.
- Yi Mao and Guy Lebanon. Sequential models for sentiment prediction. In *ICML Workshop on Learning in Structured Output Spaces*, 2006.
- Adam Marcus, Michael S Bernstein, Osama Badar, David R Karger, Samuel Madden, and Robert C Miller. Twitinfo: aggregating and visualizing microblogs for event exploration. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 227–236. ACM, 2011.

- Eugenio Martínez-Cámara, M Teresa Martín-Valdivia, L Alfonso Urena-López, and A Rturo Montejo-Ráez. Sentiment analysis in twitter. *Natural Language Engineering*, 20(01):1–28, 2014.
- Diana Maynard and Adam Funk. Automatic detection of political opinions in tweets. In *Extended Semantic Web Conference*, pages 88–99. Springer, 2011.
- Andrew McAfee, Erik Brynjolfsson, Thomas H Davenport, DJ Patil, and Dominic Barton. Big data. *The management revolution. Harvard Bus Rev*, 90(10):61–67, 2012.
- Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.
- Walaa Medhat, Ahmed Hassan, and Hoda Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113, 2014.
- Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th international conference on World Wide Web*, pages 171–180. ACM, 2007.
- Prem Melville, Wojciech Gryc, and Richard D Lawrence. Sentiment analysis of blogs by combining lexical knowledge with text classification. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1275–1284. ACM, 2009.
- Robert C Merton. A simple model of capital market equilibrium with incomplete information. *The journal of finance*, 42:483–510, 1987.
- Qingliang Miao, Qiudan Li, and Ruwei Dai. Amazing: A sentiment mining and retrieval system. *Expert Systems with Applications*, 36(3):7192–7198, 2009.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

References

- Saif Mohammad, Cody Dunne, and Bonnie Dorr. Generating High-Coverage Semantic Orientation Lexicons From Overtly Marked Words and a Thesaurus. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP '09*, pages 599–608. Association for Computational Linguistics, 2009.
- Saif M Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. *arXiv preprint arXiv:1308.6242*, 2013.
- Karo Moilanen and Stephen Pulman. Sentiment composition. In *Proceedings of RANLP*, volume 7, pages 378–382, 2007.
- Richard Montague. Formal philosophy. selected papers of richard montague. edited and with an introduction by richmond h. thomason. *New Haven*, 1974.
- Andrés Montoyo, Patricio Martínez-Barco, and Alexandra Balahur. Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments. *Decision Support Systems*, 53(4):675–679, November 2012.
- Rodrigo Moraes, João Francisco Valiati, and Wilson P Gavião Neto. Document-level sentiment classification: An empirical comparison between svm and ann. *Expert Systems with Applications*, 40(2):621–633, 2013.
- Satoshi Morinaga, Kenji Yamanishi, Kenji Tateishi, and Toshikazu Fukushima. Mining product reputations on the web. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 341–349. ACM, 2002.
- Sérgio Moro, Paulo Cortez, and Paulo Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014.
- Andrius Mudinas, Dell Zhang, and Mark Levene. 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*. ACM, 2012.

- Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. Semeval-2016 task 4: Sentiment analysis in twitter. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval 2016), San Diego, US (forthcoming)*, 2016.
- Robert Neal and Simon M Wheatley. Do Measures of Investor Sentiment Predict Returns? *Journal of Financial and Quantitative Analysis*, 33(04):523–547, 1998.
- Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. Recognition of affect, judgment, and appreciation in text. In *Proceedings of the 23rd international conference on computational linguistics*, pages 806–814. Association for Computational Linguistics, 2010.
- Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. Sentiful: A lexicon for sentiment analysis. *Affective Computing, IEEE Transactions on*, 2(1):22–36, 2011.
- Thien Hai Nguyen, Kiyooki Shirai, and Julien Velcin. Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications*, 42(24):9603–9611, 2015.
- Brendan O’Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, 11(122-129):1–2, 2010.
- Chong Oh and Olivia R Liu Sheng. Investigating Predictive Power of Stock Micro Blog Sentiment in Forecasting Future Stock Price Directional Movement. In *ICIS 2011 Proceedings*, Shanghai, China, 2011. AIS.
- Neil O’Hare, Michael Davy, Adam Bermingham, Paul Ferguson, Páraic Sheridan, Cathal Gurrin, and Alan F Smeaton. Topic-dependent sentiment analysis of financial blogs. In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, pages 9–16. ACM, 2009.

References

- Daniel E O’Leary. Blog mining-review and extensions:from each according to his opinion. *Decision Support Systems*, 51(4):821–830, 2011.
- Nuno Oliveira, Paulo Cortez, and Nelson Areal. On the predictability of stock market behavior using stocktwits sentiment and posting volume. In *Progress in Artificial Intelligence*, volume 8154 of *Lecture Notes in Computer Science*, pages 355–365. Springer Berlin Heidelberg, 2013.
- Nuno Oliveira, Paulo Cortez, and Nelson Areal. Stock market sentiment lexicon acquisition using microblogging data and statistical measures. *Decision Support Systems*, 85:62 – 73, 2016.
- Nuno Oliveira, Paulo Cortez, and Nelson Areal. The impact of microblogging data for stock market prediction: Using twitter to predict returns, volatility, trading volume and survey sentiment indices. *Expert Systems with Applications*, 73:125 – 144, 2017.
- David Omand, Jamie Bartlett, and Carl Miller. Introducing social media intelligence (socmint). *Intelligence and National Security*, 27(6):801–823, 2012.
- Iadh Ounis, Craig Macdonald, and Ian Soboroff. On the trec blog track. In *ICWSM*, 2008.
- Asil Oztekin, Recep Kizilaslan, Steven Freund, and Ali Iseri. A data analytic approach to forecasting daily stock returns in an emerging market. *European Journal of Operational Research*, 253(3):697–710, 2016.
- Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, pages 1320–1326, 2010.
- Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 115–124. Association for Computational Linguistics, 2005.

- Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, January 2008.
- Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.
- Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–43, 2014.
- Richard L Peterson. Affect and financial decision-making: How neuroscience can inform market participants. *The Journal of Behavioral Finance*, 8(2):70–78, 2007.
- Giovanni Petris. An r package for dynamic linear models. *Journal of Statistical Software*, 36(12):1–16, 2010.
- José C Pinheiro and Douglas M Bates. Unconstrained parametrizations for variance-covariance matrices. *Statistics and Computing*, 6(3):289–296, 1996.
- Robi Polikar. Ensemble based systems in decision making. *IEEE Circuits and systems magazine*, 6(3):21–45, 2006.
- Soujanya Poria, Erik Cambria, Gregoire Winterstein, and Guang-Bin Huang. Sentic patterns: Dependency-based rules for concept-level sentiment analysis. *Knowledge-Based Systems*, 69:45–63, 2014.
- Rudy Prabowo and Mike Thelwall. Sentiment analysis: A combined approach. *Journal of Informetrics*, 3(2):143–157, 2009.
- Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Opinion word expansion and target extraction through double propagation. *Computational Linguistics*, 37(1):9–27, 2011.
- Lily Qiu and Ivo Welch. Investor Sentiment Measures. *Seminar*, 10794:1–39, 2006.

References

- R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL <http://www.R-project.org/>.
- Lior Rokach. Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1-2):1–39, 2010.
- Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.
- Sara Rosenthal, Alan Ritter, Preslav Nakov, and Veselin Stoyanov. Semeval-2014 task 9: Sentiment analysis in twitter. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 73–80, 2014.
- Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif M Mohammad, Alan Ritter, and Veselin Stoyanov. Semeval-2015 task 10: Sentiment analysis in twitter. *Proceedings of SemEval-2015*, 2015a.
- Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif M Mohammad, Alan Ritter, and Veselin Stoyanov. Semeval-2015 task 10: Sentiment analysis in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval*, 2015b.
- Huaxia Rui, Yizao Liu, and Andrew Whinston. Whose and what chatter matters? the effect of tweets on movie sales. *Decision Support Systems*, 55(4):863–870, 2013.
- David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, DTIC Document, 1985.
- Philip Russom. Big data analytics. *TDWI Best Practices Report, Fourth Quarter*, pages 1–35, 2011.
- Sanjiv Sabherwal, Salil K Sarkar, and Ying Zhang. Do internet stock message boards influence trading? evidence from heavily discussed stocks with no fundamental news. *Journal of Business Finance & Accounting*, 38(9-10):1209–1237, 2011.
- Hassan Saif, Yulan He, and Harith Alani. Semantic sentiment analysis of twitter. In *The Semantic Web–ISWC 2012*, pages 508–524. Springer, 2012.

- Maik Schmeling. Institutional and individual sentiment: Smart money and noise trader risk? *International Journal of Forecasting*, 23(1):127–145, 2007.
- Maik Schmeling. Investor sentiment and stock returns: Some international evidence. *Journal of Empirical Finance*, 16(3):394–408, June 2009.
- Robert P Schumaker and Hsinchun Chen. Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Transactions on Information Systems (TOIS)*, 27(2):12, 2009.
- Robert P. Schumaker, Yulei Zhang, Chun-Neng Huang, and Hsinchun Chen. Evaluating sentiment in financial news articles. *Decision Support Systems*, 53(3):458–464, June 2012.
- Giovanni Seni and John F Elder. Ensemble methods in data mining: improving accuracy through combining predictions. *Synthesis Lectures on Data Mining and Knowledge Discovery*, 2(1):1–126, 2010.
- Catherine T Shalen. Volume, volatility, and the dispersion of beliefs. *Review of Financial Studies*, 6(2):405–434, 1993.
- Her-Jiun Sheu and Yu-Chen Wei. Effective options trading strategies based on volatility forecasting recruiting investor sentiment. *Expert Systems with Applications*, 38(1):585–596, 2011.
- Robert J Shiller. From efficient markets theory to behavioral finance. *The Journal of Economic Perspectives*, 17(1):83–104, 2003.
- Andrei Shleifer and Robert W Vishny. The limits of arbitrage. *The Journal of Finance*, 52(1):35–55, 1997.
- Herbert A Simon. *The sciences of the artificial*. MIT press, 1996.

References

- Jasmina Smailović, Miha Grčar, Nada Lavrač, and Martin Žnidaršič. Predictive sentiment analysis of tweets: A stock market application. In *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data*, pages 77–88. Springer, 2013.
- Jasmina Smailović, Miha Grčar, Nada Lavrač, and Martin Žnidaršič. Stream-based active learning for sentiment analysis in the financial domain. *Information Sciences*, 285:181–203, 2014.
- Alex J Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004.
- Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, volume 1631. Citeseer, 2013.
- Michael E. Solt and Meir Statman. How Useful is the Sentiment Index? *Financial Analysts Journal*, 44(5):45–55, 1988.
- Ellen Spertus. Smokey: Automatic recognition of hostile messages. In *AAAI/IAAI*, pages 1058–1065, 1997.
- Timm O Sprenger, Andranik Tumasjan, Philipp G Sandner, and Isabell M Welp. Tweets and trades: the information content of stock microblogs. *European Financial Management*, 20(5):926–957, 2014.
- Robert F. Stambaugh, Jianfeng Yu, and Yu Yuan. The short of it: Investor sentiment and anomalies. *Journal of Financial Economics*, 104(2):288–302, May 2012.
- Philip Stone, Dexter Dunphy, Marshall Smith, and Daniel Ogilvie. *The General Inquirer: A Computer Approach to Content Analysis*, volume 08. MIT Press, 1966.

- Veselin Stoyanov, Claire Cardie, and Janyce Wiebe. Multi-perspective question answering using the opqa corpus. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 923–930. Association for Computational Linguistics, 2005.
- Maite Taboada, Caroline Anthony, and Kimberly Voll. Methods for creating semantic orientation dictionaries. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LRECo6)*, pages 427–432, 2006.
- Hiroya Takamura, Takashi Inui, and Manabu Okumura. Extracting semantic orientations of words using spin model. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 133–140. Association for Computational Linguistics, 2005.
- Duyu Tang, Furu Wei, Bing Qin, Ting Liu, and Ming Zhou. Coooolll: A deep learning system for twitter sentiment classification. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 208–212, 2014a.
- Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. Learning sentiment-specific word embedding for twitter sentiment classification. In *ACL (1)*, pages 1555–1565, 2014b.
- Jie Tang, Jimeng Sun, Chi Wang, and Zi Yang. Social influence analysis in large-scale networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 807–816. ACM, 2009.
- Leonard J Tashman. Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting*, 16(4):437–450, 2000.
- Loren Terveen, Will Hill, Brian Amento, David McDonald, and Josh Creter. Phoaks: A system for sharing recommendations. *Communications of the ACM*, 40(3):59–62, 1997.
- Paul C Tetlock. Giving Content to Investor Sentiment : The Role of Media in the Stock Market. *Journal of Finance*, 62(3):1139–1168, 2007.

References

- Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. Sentiment in twitter events. *Journal of the American Society for Information Science and Technology*, 62(2):406–418, 2011.
- Lyn C Thomas. A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International journal of forecasting*, 16(2):149–172, 2000.
- Matt Thomas, Bo Pang, and Lillian Lee. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 327–335. Association for Computational Linguistics, 2006.
- Allan Timmermann. Elusive return predictability. *International Journal of Forecasting*, 24(1):1–18, 2008.
- Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology NAACL 03*, 1(June):173–180, 2003.
- Dennis Tsichritzis. The dynamics of innovation. In *Beyond calculation*, pages 259–265. Springer, 1997.
- Mikalai Tsytsarau and Themis Palpanas. Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery*, 24(3):478–514, 2012.
- Dan Tufiş and Dan Ştefănescu. Experiments with a differential semantics annotation for wordnet 3.0. *Decision Support Systems*, 53(4):695–703, 2012.
- Robert Tumarkin and Robert F Whitelaw. News or noise? internet postings and stock prices. *Financial Analysts Journal*, 57(3):41–51, 2001.

- Andranik Tumasjan, Timm O Sprenger, Philipp G Sandner, and Isabell M Welp. Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, 10:178–185, 2010.
- Efraim Turban, Ramesh Sharda, Jay E Aronson, and David King. *Business intelligence: A managerial approach*. Prentice Hall, 2008.
- Peter D. Turney. 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*, pages 417–424. Association for Computational Linguistics, 2002.
- Peter D. Turney and Michael L. Littman. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Trans. Inf. Syst.*, 21(4):315–346, October 2003.
- Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., 1995.
- Vladimir Vapnik and Alexey Chervonenkis. A note on one class of perceptrons. *Automation and remote control*, 25(1), 1964.
- Vladimir Vapnik and Alexey Chervonenkis. *Theory of Pattern Recognition [in Russian]*. Nauka, Moscow, 1974.
- Vladimir Vapnik and A. Lerner. Pattern recognition using generalized portrait method. *Automation and remote control*, 24, 1963.
- Rahul Verma and Gökçe Soydemir. The impact of individual and institutional investor sentiment on the market price of risk. *The Quarterly Review of Economics and Finance*, 49(3):1129–1145, 2009.

References

- Rahul Verma and Priti Verma. Are survey forecasts of individual and institutional investor sentiments rational? *International Review of Financial Analysis*, 17(5):1139–1155, 2008.
- Iris Vessey and Robert Glass. Strong vs. weak approaches to systems development. *Communications of the ACM*, 41(4):99–102, 1998.
- Gang Wang, Jinxing Hao, Jian Ma, and Hongbing Jiang. A comparative assessment of ensemble learning for credit scoring. *Expert systems with applications*, 38(1):223–230, 2011.
- Greg Welch and Gary Bishop. An introduction to the kalman filter. Technical report, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, 1995.
- Janyce Wiebe. Learning subjective adjectives from corpora. In *AAAI/IAAI*, pages 735–740, 2000.
- Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. OpinionFinder : A system for subjectivity analysis. *October*, pages 34–35, 2005.
- Ian H Witten, Eibe Frank, and Mark A Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers Inc., 2011.
- Peter D Wysocki. Cheap talk on the web: The determinants of postings on stock message boards. *University of Michigan Business School Working Paper*, 1998.
- Kaiquan Xu, Stephen Shaoyi Liao, Jiexun Li, and Yuxia Song. Mining comparative opinions from customer reviews for competitive intelligence. *Decision support systems*, 50(4):743–754, 2011.
- Jeonghee Yi, Tetsuya Nasukawa, Razvan Bunescu, and Wayne Niblack. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing

- techniques. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pages 427–434. IEEE, 2003.
- Paul D Yoo, Maria H Kim, and Tony Jan. Machine learning techniques and use of event information for stock market prediction: A survey and evaluation. In *International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06)*, volume 2, pages 835–841. IEEE, 2005.
- Hong Yu and Vasileios Hatzivassiloglou. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 129–136. Association for Computational Linguistics, 2003.
- Jianfeng Yu and Yu Yuan. Investor sentiment and the mean-variance relation. *Journal of Financial Economics*, 100(2):367–381, 2011.
- Yang Yu, Wenjing Duan, and Qing Cao. The impact of social and conventional media on firm equity value: A sentiment analysis approach. *Decision Support Systems*, 55(4): 919–926, November 2013.
- Dongsong Zhang and Lina Zhou. Discovering golden nuggets: data mining in financial application. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 34(4):513–522, 2004.
- Lei Zhang, Riddhiman Ghosh, Mohamed Dekhil, Meichun Hsu, and Bing Liu. Combining lexicon-based and learning-based methods for twitter sentiment analysis. *HP Laboratories Technical Report*, 2011a.
- Xue Zhang, Hauke Fuehres, and Peter A Gloor. Predicting stock market indicators through twitter i hope it is not as bad as i fear. *Procedia-Social and Behavioral Sciences*, 26:55–62, 2011b.

References

- Zhaohui Zheng, Xiaoyun Wu, and Rohini Srihari. Feature selection for text categorization on imbalanced data, 2004.
- Cäcilia Zirn, Mathias Niepert, Heiner Stuckenschmidt, and Michael Strube. Fine-grained sentiment analysis with structural features. In *IJCNLP*, pages 336–344, 2011.
- Mohamed Zouaoui, Genevieve Nouyrigat, and Francisca Beer. How does investor sentiment affect stock market crises? evidence from panel data. *Financial Review*, 46(4):723–747, 2011.

INDEX

- Alm et al. (2005), 54
- Al Nasser et al. (2015), 76, 82, 122
- Amit and Geman (1997), 42, 43
- Annett and Kondrak (2008), 52, 54
- Antweiler and Frank (2004), 4, 5, 73, 76, 77, 82, 84, 92, 122, 123, 128, 129, 134, 151, 165, 166
- Areal and Taylor (2002), 135
- Armstrong and Collopy (1992), 140
- Armstrong (2001), 140
- Asur and Huberman (2010), 54, 66
- Atsalakis and Valavanis (2009), 72, 73
- Baccianella et al. (2010), 51, 57, 60, 62, 92, 93, 102, 115
- Bai (2011), 54
- Baker and Wurgler (2006), 75, 76, 79, 131, 144, 145, 148, 156, 158–160, 170
- Baker and Wurgler (2007), 73, 75
- Baker et al. (2012), 76, 84, 122, 131, 144, 145, 148, 156, 158–160, 170
- Birmingham and Smeaton (2011), 54
- Biau (2012), 43
- Bifet and Frank (2010), 54
- Bishop (2006), xv, 30–35
- Bollen et al. (2011), 4, 5, 54, 66, 73, 76–78, 80–82, 92, 120–122, 134, 141, 156, 159, 165, 166, 170
- Boser et al. (1992), 38
- Breiman (2001), 41–44
- Brown and Cliff (2004), 75–77, 79, 81, 84, 104, 106, 114, 120, 122, 123, 129
- Brown and Cliff (2005), 76, 77, 84, 122
- Burges (1998), xv, 37, 41
- Cambria et al. (2012), 53, 54, 66
- Cambria et al. (2013), 53
- Chaovalit and Zhou (2005), 52, 54
- Chen and Lazer (2013), 5, 76, 77, 81, 121, 166
- Chen and Zhang (2014), 3, 69
- Chen et al. (2014), 3, 4
- Choi and Cardie (2008), 53, 54
- Choi et al. (2009), 54, 59
- Choi et al. (2016), 3
- Chung et al. (2012), 75
- Corredor et al. (2013), 75–77
- Cortes and Vapnik (1995), 38

Index

- Cortez (2010), 137
- Crone et al. (2011), 140
- Cybenko (1989), 32
- Da et al. (2011), 73
- Dang and Owczarzak (2008), 59
- Das and Chen (2007), 4, 5, 76, 77, 84, 122, 165, 166
- Das et al. (2005), 5, 76, 77, 166
- Dave et al. (2003), 54, 65
- Deng et al. (2011), 76, 77, 80, 121, 134, 137, 141
- Denning (1997), 8, 9
- Devitt and Ahmad (2007), 54
- Diebold and Mariano (1995), xxv, 11, 142
- Dietterich (2000), 41, 42, 45
- Duric and Song (2012), 54
- Ekman et al. (2013), 49
- Enke and Thawornwong (2005), 72, 73
- Esuli and Sebastiani (2005), 60, 62
- Fama (1998), 5, 122
- Fan and Gordon (2014), 3, 4, 47, 119
- Feldman (2013), 59, 91
- Fersini et al. (2014), 91
- Fisher and Statman (2000), 75–77, 104, 129, 138, 158
- Garcia (2013), 5, 76, 77, 80, 166
- Genuer et al. (2008), 43
- Geva and Zahavi (2014), 77
- Gindl and Liegl (2008), 52
- Go et al. (2009), 54
- Godbole et al. (2007), 54
- Goldberg and Zhu (2006), 54
- Goldberg et al. (2001), 141
- Greene (2000), 28
- Groß-Klußmann and Hautsch (2011), 5, 76, 77, 166
- Gundecha and Liu (2012), 4, 160
- Hagenau et al. (2013), 54, 66, 76, 77
- Han and Lee (2016), 4
- Han et al. (2011), 25, 26, 33–35, 39–41, 43, 44
- Hansen and Salamon (1990), 41, 45
- Hastie et al. (2008), 135
- Hastie et al. (2013), xv, 26, 28, 29
- Hatzivassiloglou and McKeown (1997), 57, 59, 62, 63
- Hatzivassiloglou and Wiebe (2000), 59
- He and Zhou (2011), 54
- Heerschop et al. (2011), 53, 54
- Hengelbrock et al. (2013), 77
- Hevner and Chatterjee (2010), 8
- Hevner et al. (2004), 8, 9
- Hirshleifer and Teoh (2003), 74
- Ho and Hung (2009), 75–77
- Ho (1998), 42
- Hribar and McInnis (2012), 75
- Hu and Liu (2004), 54, 57, 59, 61–63, 65, 91, 93, 102

- Hu et al. (2012), 54
- Huang et al. (2004), 71
- Hutto and Gilbert (2014), 54
- Hyndman and Koehler (2006), 140, 141
- Jiang et al. (2011), 54
- Kamps et al. (2004), 60, 62, 63
- Kang et al. (2012), 54, 65
- Kennedy and Inkpen (2006), 59, 62
- Kim and Hovy (2004), 54, 60, 62
- Kiritchenko et al. (2014), 51, 54, 60, 62, 63, 91, 94, 98, 101
- Kohavi (1995), xxv
- Kontopoulos et al. (2013), 53, 54
- Ku et al. (2006), 54
- Kudo and Matsumoto (2004), 54, 57
- Kumar and Ravi (2007), 70
- Kurov (2008), 75, 77
- Kurov (2010), 75–77
- Kwon et al. (2006), 66
- Lampos et al. (2013), 54, 66
- Laney (2001), 3, 69
- Laver et al. (2003), 54, 66
- Lee et al. (1991), 76
- Lee et al. (2002), 76, 77, 144
- Lemmon and Portniaguina (2006), 75
- Lerman et al. (2009), 54
- Leung et al. (2006), 54
- Lewis (1992), 97
- Li and Li (2013), 54
- Li et al. (2014), 77
- Lin and He (2009), 54
- Lin et al. (2012), 70
- Lin (1998), 59
- Lita et al. (2005), 66
- Liu and Zhang (2012), 47–50
- Liu et al. (2003), 54
- Long et al. (1990), 73, 74, 148
- Loughran and McDonald (2011), 59, 62, 78, 92–94, 102, 115, 120
- Lu et al. (2011), 61
- Makrehchi et al. (2013), 77
- Mao and Lebanon (2006), 58
- Mao et al. (2011), 76, 78, 81, 120, 121, 141
- Mao et al. (2014), 62, 92, 94, 115
- Marcus et al. (2011), 54
- Martínez-Cámara et al. (2014), 52, 64
- Maynard and Funk (2011), 54
- McAfee et al. (2012), 3
- McCulloch and Pitts (1943), 29
- Medhat et al. (2014), 52
- Mei et al. (2007), 54
- Melville et al. (2009), 54
- Merton (1987), 74
- Miao et al. (2009), 54
- Mikolov et al. (2013), 58
- Mohammad et al. (2009), 60, 62, 93, 102
- Mohammad et al. (2013), 51, 54
- Moilanen and Pulman (2007), 53, 54

Index

- Montague (1974), 53
- Montoyo et al. (2012), 91
- Moraes et al. (2013), 54
- Morinaga et al. (2002), 54
- Moro et al. (2014), xxvi, 94
- Mudinas et al. (2012), 53, 54, 65
- Nakov et al. (2016), 53, 55, 57, 58
- Neal and Wheatley (1998), 76, 84, 122, 144
- Neviarouskaya et al. (2010), 53, 54
- Neviarouskaya et al. (2011), 61–63
- Nguyen et al. (2015), 76, 80–82, 121, 122
- O'Connor et al. (2010), 54, 66
- O'Hare et al. (2009), 54, 66
- O'Leary (2011), 63, 109
- Oh and Sheng (2011), 4, 5, 76, 77, 80, 81, 121, 137, 165, 166
- Oliveira et al. (2013), 128, 134, 151
- Oliveira et al. (2016), 18, 124, 126, 157
- Oliveira et al. (2017), 18
- Omand et al. (2012), 4
- Ounis et al. (2008), 59
- Oztekin et al. (2016), 85, 136
- Pak and Paroubek (2010), 54
- Pang and Lee (2005), 54, 66
- Pang and Lee (2008), xxvi, 47, 57, 85
- Pang et al. (2002), 57
- Pennington et al. (2014), 58
- Peterson (2007), 72
- Petris (2010), 131
- Pinheiro and Bates (1996), 131
- Polikar (2006), 44–46
- Poria et al. (2014), 53, 54
- Prabowo and Thelwall (2009), 52, 54
- Qiu and Welch (2006), 75, 76
- Qiu et al. (2011), 60, 62, 63, 94
- R Core Team (2013), 95
- Rokach (2010), 44–46
- Rosenblatt (1958), 29
- Rosenthal et al. (2014), 53, 55, 57, 64
- Rosenthal et al. (2015a), 53, 55, 57, 58
- Rosenthal et al. (2015b), 59, 91
- Rui et al. (2013), 54
- Rumelhart et al. (1985), 29, 32
- Russom (2011), 3
- Sabherwal et al. (2011), 5, 76, 77, 82, 122, 134, 151, 156, 159, 166, 170
- Saif et al. (2012), 91
- Schmeling (2007), 76, 77, 81, 122, 129
- Schmeling (2009), 75, 76
- Schumaker and Chen (2009), 92
- Schumaker et al. (2012), 5, 76, 77, 80, 82, 92, 121, 122, 166
- Seni and Elder (2010), xv, 41, 42
- Shalen (1993), 129, 151
- Sheu and Wei (2011), 76
- Shiller (2003), 4, 73
- Shleifer and Vishny (1997), 74

- Simon (1996), 13
- Smailović et al. (2013), 76, 77
- Smailović et al. (2014), 77
- Smola and Schölkopf (2004), xv, 39, 40
- Socher et al. (2013), 53, 54
- Solt and Statman (1988), 76, 77, 129, 138, 158
- Spertus (1997), 54, 66
- Sprenger et al. (2014), 4, 5, 76, 77, 84, 122, 128, 134, 151, 156, 159, 165, 166, 170
- Stambaugh et al. (2012), 74–76
- Stone et al. (1966), 51, 57, 59, 92, 93, 102, 115
- Stoyanov et al. (2005), 66
- Taboada et al. (2006), 54
- Takamura et al. (2005), 61–63
- Tang et al. (2009), 4
- Tang et al. (2014a), 54
- Tang et al. (2014b), 54
- Tashman (2000), 141
- Terveen et al. (1997), 66
- Tetlock (2007), 5, 76, 77, 80, 82, 122, 128, 134, 137, 151, 166
- Thelwall et al. (2011), 54
- Thomas et al. (2006), 54, 66
- Thomas (2000), 71
- Timmermann (2008), 5, 84, 123, 166
- Toutanova et al. (2003), 95, 127
- Tsichritzis (1997), 8, 9
- Tsytsarau and Palpanas (2012), 48, 49, 51, 52
- Tufiş and Ştefănescu (2012), 62
- Tumarkin and Whitelaw (2001), 5, 76, 77, 84, 123, 166
- Tumasjan et al. (2010), 54
- Turban et al. (2008), 25, 26, 41
- Turney and Littman (2003), 59, 60, 62, 96, 98
- Turney (2002), 65, 66
- Vapnik and Chervonenkis (1964), 36
- Vapnik and Chervonenkis (1974), 36
- Vapnik and Lerner (1963), 36
- Vapnik (1995), 39
- Verma and Soydemir (2009), 75, 77, 104, 106, 129, 131, 138, 158
- Verma and Verma (2008), 75, 77
- Vessey and Glass (1998), 13
- Wang et al. (2011), 71
- Welch and Bishop (1995), xxv
- Wiebe (2000), 57, 59, 62, 63
- Wilson et al. (2005), 51, 57, 59, 92, 93, 102, 115
- Witten et al. (2011), 22, 25, 32, 34–36, 39, 40
- Wysocki (1998), 5, 134, 151, 166
- Xu et al. (2011), 54
- Yi et al. (2003), 54

Index

Yoo et al. (2005), 73

Yu and Hatzivassiloglou (2003), 54

Yu and Yuan (2011), 75, 76

Yu et al. (2013), 76, 77, 92

Zhang and Zhou (2004), 72, 73

Zhang et al. (2011a), 54

Zhang et al. (2011b), 76, 77

Zheng et al. (2004), 97

Zirn et al. (2011), 54

Zouaoui et al. (2011), 75

da Silva et al. (2014), 91

Part IV

APPENDICES



CODE

A.1 SHORT EXAMPLES OF THE DEVELOPED R CODE

This appendix contains some short examples of R code applied in this thesis. The first example presents the execution of authorized requests to the Twitter REST API to collect tweets about specific cashtags with a minimum tweet *ID*. This project did not apply specialized R packages (e.g., *twitterR*) to get Twitter data because these packages only allow the collection of a small set of attributes and this work needs a higher number of attributes. The collection of Twitter data about all stocks traded in US markets applies this request sequentially for groups of stocks because there are rate limits and parameters constraints. For instance, it is not possible to include all cashtags in the parameters and REST API permits to collect a maximum of 100 tweets in each request.

Twitter uses OAuth to provide authorized access to its API. To construct the HTTP request, the function *reqTweet* needs to use the application settings that are available at the settings page for your application on dev.twitter.com/apps. Moreover, it is necessary to create a OAuth 1.0a HMAC-SHA1 signature (*oauth_signature*), an unique token for the request (*nonce*) and a timestamp (*oauth_timestamp*). The signature allows Twitter to verify that the request has not been changed in transit, verify the application issuing the request, and verify that the application is authorized to interact with the

Appendix A. Code

users account. The function `reqTweet` collects a maximum of 100 tweets that are stored by the function `saveTweets`. This function processes the results provided by Twitter REST API in the JSON format and saves various attributes (e.g., text, id, created time) in a MongoDB database. This appendix does not include function `saveTweets` because it is very long.

```
library(base64enc) # loads package base64enc: provides tools for handling base64 encoding
library(RCurl) # loads package RCurl: provides functions to compose HTTP requests, fetch
  URIs, get & post forms and process the returned results
library(digest) # loads package digest: create compact hash digests of R objects

### application settings ###
consumer_key<-<consumer key>
access_token<-<access token>
consumer_secret<-<consumer secret>
access_token_secret<-<access token secret>

### gen_nonce: generates an unique token for each unique request ###
gen_nonce <- function() {
  letters<-"ABCDEFGHIJKLMNOPQRSTUVWXYZabcdefghijklmnopqrstuvwxyz"
  letters<-unlist(strsplit(letters,"")) # creates a vector of alphabetic characters
  nonce<-paste(replicate(32,sample(letters,1)),collapse="") # produces a string of 32
  alphabetic characters
  nonce<-tolower(base64(nonce)) # base64 encoding of string
  return (substr(nonce,1,32)) # returns 32 bytes of random data
}

### reqTweet: executes authorized requests to REST API. The inputs are the vector of
  tickers to be collected (tickers) and the minimum id (min_id) ###
reqTweet <- function(tickers, min_id) {
  oauth_signature<-"+"
  # creation of a valid OAuth 1.0a HMAC-SHA1 signature
  while (grepl("+",oauth_signature,fixed=TRUE,useBytes=TRUE)) {
    nonce<-gen_nonce(); # generates an unique token for this request
    oauth_timestamp<-as.integer(Sys.time()) # creates a timestamp: number of seconds since
    the Unix epoch
    query<-paste0("q%3D%2524",paste(toupper(tickers),collapse="%2520OR%2520%2524"),"
    %2520%26since_id%3D",min_id) # generates a percent encoded parameter string to
    collect tweets having cashtags about the provided tickers
```

```

signbase<-paste0("GET&https%3A%2F%2Fapi.twitter.com%2F1.1%2Fsearch%2Ftweets.json&count
%3D100%26oauth_consumer_key%3D",consumer_key,"%26oauth_nonce%3D",nonce,"%26oauth_
signature_method%3DHMAC-SHA1%26oauth_timestamp%3D",oauth_timestamp,"%26oauth_token
%3D",access_token,"%26oauth_version%3D1.0%26",query) # creates the signature base
string to collect 100 tweets containing cashtags about the provided tickers with
an ID greater than min_id

signkey<-paste0(consumer_secret,"&",access_token_secret) # creates the signing key:
percent encoded consumer secret, followed by the character & , followed by
the percent encoded token secret

oauth_signature<-URLencode(base64(hmac(key=signkey, object=signbase, algo="sha1", raw=
TRUE)),reserved=TRUE) # generates a OAuth 1.0a HMAC-SHA1 signature
}

headr<-paste0('Authorization: OAuth oauth_consumer_key=',consumer_key,', oauth_nonce='
',nonce,', oauth_signature=',oauth_signature,', oauth_signature_method="HMAC-SHA1
", oauth_timestamp=',oauth_timestamp,', oauth_token=',access_token,', oauth_
version="1.0') # builds the header string

url_text<-paste0("https://api.twitter.com/1.1/search/tweets.json?count=100&q=%24",paste(
toupper(tickers),collapse="+OR+%24"),"&since_id=",min_id) # creates URL address

req_json<-try(getURL(url=url_text, httpheader=headr, ssl.verifypeer = FALSE)) # makes
API request and retrieves data

if (is.null(attr(req_json,"condition"))) saveTweets(req_json) # stores data if the API
call was valid

}

```

Listing A.1: Twitter Data Collection

The second example presents two functions applied in the production of an unique daily sentiment indicator from daily microblogging indicators and weekly and monthly survey measures using a Kalman Filter procedure. The *buildModel* function generates a dynamic linear model that is used in the maximum likelihood estimation of the parameters. To diminish the complexity of the optimization problem and assure that the variance-covariance matrix is positive semi-definite, this function applies the approach proposed by Pinheiro and Bates (1996) and followed by Petris (2010), parametrizing the covariance matrix (V) in terms of the elements of its log-Cholesky decomposition and the system variance (W) using its log. The *kfilter* function receives a data frame

Appendix A. Code

containing the sentiment values of various sources and the final training date. Table 23 shows 30 days of data frame *sent*. To create the Kalman Filter indicator, the *kfilter* function estimates the parameters using maximum likelihood for the training period, fits the model using the estimated parameters and *buildModel* function and filters the whole series using the fitted model.

```
library(dlm) # loads package dlm, for bayesian and likelihood analysis of dynamic linear
             models

### buildModel: creates a dynamic linear model that is applied in the maximum likelihood
             estimation of the parameters ###
buildModel <- function(x){
  L <- matrix(0, 5, 5)
  L[upper.tri(L, TRUE)] <- x[1 : 15]
  diag(L) <- exp(diag(L))
  mF <- matrix(1, 5, 1)
  b_model <- dlm(FF = mF, V = crossprod(L),
                GG = 1, W = exp(x[16]), m0 = 0, CO = 1e7) # creates the dynamic linear model: covariance
                matrix (V) is parametrized in terms of the elements of its log-Cholesky
                decomposition and the system variance (W) uses its log.
  return(b_model) # returns the created dynamic linear model
}

### kfilter: produces an unique sentiment indicator from various sentiment measures using
             the Kalman Filter procedure. This function takes the following inputs: sent is a data
             frame containing the sentiment values of diverse sources and respective date; train_
             day is the last training day ###
kfilter <- function (sent, train_day) {
  sent_train<-as.matrix(sent[sent$Date<=train_day,colnames(sent)!="Date"]) # creates a
  training data set containing data until the last training day
  y<-as.matrix(sent[,colnames(sent)!="Date"]) # creates a matrix containing the sentiment
  value of all sources for the entire time period
  mparam <- rep(0, 16)
  fitModel <- dlmMLE(sent_train, parm = mparam, build = buildModel,
                    hessian = TRUE, control = list(maxit = 1000)) # estimates the parameters by maximum
                    likelihood
  estimated_model <- buildModel(fitModel$par) # fits the model using the build_model
  function and the estimated parameters
  filteredseries <- dlmFilter(y, estimated_model) # calculates the filtered values for the
  entire series using the fitted model
}
```


Table 23.: Sample (30 days) of the data frame *sent* containing sentiment values of daily Twitter indicator, weekly survey indicators (AAII and II) and monthly survey indicators (UMCS and Sentix)

Date	AAII	II	UMCS	Sentix	Twitter
2012-12-22	NA	NA	NA	NA	-1.95145381
2012-12-23	NA	NA	NA	NA	-2.92451397
2012-12-24	NA	NA	NA	NA	-3.15813721
2012-12-25	NA	NA	NA	NA	-2.14482210
2012-12-26	NA	NA	NA	NA	-2.65627487
2012-12-27	0.3365247	-0.72547738	NA	NA	-3.18382091
2012-12-28	NA	NA	NA	NA	-2.91555930
2012-12-29	NA	NA	NA	NA	-2.67729967
2012-12-30	NA	NA	NA	NA	-5.33736293
2012-12-31	NA	NA	NA	NA	-2.31824411
2013-01-01	NA	NA	NA	NA	-2.86469216
2013-01-02	NA	NA	NA	NA	-1.95909316
2013-01-03	-0.5857993	-0.32589109	NA	NA	-1.74535991
2013-01-04	NA	NA	NA	NA	-2.48194411
2013-01-05	NA	NA	NA	NA	-1.01733527
2013-01-06	NA	NA	NA	NA	-2.04312003
2013-01-07	NA	NA	NA	NA	-1.75028230
2013-01-08	NA	NA	NA	NA	-2.57602862
2013-01-09	NA	-0.03528288	NA	NA	-2.20421822
2013-01-10	0.7593553	NA	NA	NA	-1.35432549
2013-01-11	NA	NA	NA	NA	-0.08552213
2013-01-12	NA	NA	NA	NA	-1.11403080
2013-01-13	NA	NA	NA	NA	-0.69678335
2013-01-14	NA	NA	NA	NA	-0.16392442
2013-01-15	NA	NA	-1.644492	-2.198326	-0.07654753
2013-01-16	NA	-0.03528288	NA	NA	-0.26488735
2013-01-17	0.5293671	NA	NA	NA	-0.55481751
2013-01-18	NA	NA	NA	NA	-0.36457293
2013-01-19	NA	NA	NA	NA	-1.39758923
2013-01-20	NA	NA	NA	NA	-1.03514280

```
return(filteredseries$m[2:length(filteredseries$m)]) # returns the filtered values
}
```

Listing A.2: Kalman Filter Procedure

The third example shows a simple application of the pairwise Diebold-Mariano statistical test for predictive accuracy. The function takes three different vectors: *forecast1* is a vector of predicted values produced by a predictive model, *forecast2* is a vector of forecasts generated by a baseline model and *observ* is the vector of observed values. It calculates the errors of both forecasting models and uses function *dm.test* of package *forecast* to apply the Diebold-Mariano test on these error vectors. The defined alternative hypothesis is that model 2 (baseline) is less accurate than model 1.

```
library(forecast) # loads package forecast: contains forecasting functions for time series
and linear models
### DM_test: application of the pairwise Diebold-Mariano statistical test of predictive
accuracy. This function takes the following inputs: forecast1 is a vector of predicted
```

Appendix A. Code

```
values computed by a predictive model; forecast2 is a vector of predicted values
computed by a baseline model; observ is the vector of the observed values ###
DM_test <- function (forecast1, forecast2, observ) {
# the execution of this function stops if vectors do not have the same length
  if (length(forecast1) != length(observ) || length(forecast2) != length(observ)) {
    stop("Vectors must have the same length.")
  }
  error1 <- observ - forecast1 # creates the vector of errors made by model 1
  error2 <- observ - forecast2 # creates the vector of errors made by model 2 (baseline)
  v_test<-dm.test(error1, error2, alternative="less", power=1) # applies the pairwise
  Diebold-Mariano test for both models: alternative hypothesis is that model 2 (
  baseline) is less accurate than model 1
  return(v_test) # returns the values produced by the Diebold-Mariano test
}
```

Listing A.3: Application of Diebold-Mariano Test

DATA

B.1 SHORT EXAMPLE OF THE PROPOSED LEXICON

This section presents a short sample of a microblogging stock market lexicon created by the proposed procedure to automatically produce a lexicon using diverse statistical measures and a large set of labeled messages from StockTwits. This lexicon is adapted to stock market conversations in microblogging services (e.g., StockTwits, Twitter). Table 24 shows 45 entries of the microblogging stock market lexicon. The attributes of this lexical resource are:

- Item: lexical item, either an unigram or a bigram.
- POS: POS tag based on the Penn Treebank POS tagset (e.g., JJ - Adjective, NN - Noun, RB - Adverb)
- Aff_Score: Sentiment score in affirmative (i.e., non-negated) contexts.
- Neg_Score: Sentiment score in negated contexts.

B.2 SHORT EXAMPLES OF THE ANALYZED DATA

This section presents an example of a tweet containing the Apple cashtag (i.e., \$AAPL) collected from Twitter REST API using R tool. Listing B.1 shows the truncated JSON

Appendix B. Data

Table 24.: Short sample (45 entries) of the created microblogging stock market lexicon

Item	POS	Aff_Score	Neg_Score
wedge breakdown		-10.509000	-6.9660000
short opportunity		-10.468063	-6.3097297
short entry		-9.321164	-5.9445217
short candidate		-8.487323	-5.6200615
buy area		7.811000	7.8520000
bearish signal		-7.144000	-5.9899048
flow decline		-7.042000	-7.0020000
nice volume		6.833790	4.6146970
watch over		6.804906	6.8497605
buy point		6.795971	4.5788108
bearish price		-6.772000	-6.7320000
short watch		-6.721244	-5.9773043
daily bearish		-6.685000	-6.6440000
short break		-6.685000	-6.6440000
bearish	JJ	-6.633736	-0.7936923
long setup		6.451282	6.1201671
lead negative		-6.439000	-6.5020000
top form		-6.328000	-6.2880000
avoid for		-6.270000	-6.288000
let drop		-6.243263	-6.204316
list over		6.198000	6.239000
bullish macd		6.180000	6.220000
rest here		6.161000	6.201000
enter short		-6.160000	-6.139368
bearish setup		-6.144000	-5.104000
island top		-6.144000	-6.229000
NUMc negative		-6.144000	-6.104000
short trigger		-6.144000	-6.168000
remain below		-6.113250	-6.106341
fall hard		-6.077000	-6.104000
bearflag	NN	-5.836000	-5.8140000
b/o	NN	5.559534	0.3002222
short	NN	-4.837838	-4.8271946
downside	NN	-4.834469	-0.0540000
strong	RB	4.601000	4.6230000
shortable	JJ	-4.586214	-4.6235172
vulnerable	JJ	-4.438133	-4.4818182
ominous	JJ	-4.292720	-4.2724800
bullflag	NN	4.264000	4.2870000
reshort	VB	-4.128182	-4.2218182
breakdown	NN	-3.990916	-3.9125782
sell	NN	-3.832246	0.9712258
deflation	NN	-3.809091	-3.7890909
toxic	JJ	-3.750909	-3.7309091
downward	NN	-3.750882	-3.7314706

text returned by the API and Table 25 presents some tweet information converted from JSON format. Personal details were excluded for privacy reasons.

```
{
  "statuses": [
    {
      "created_at": "Wed Jan 18 13:15:30 +0000 2017",
      "id": 821707589226352640,
      "id_str": "821707589226352640",
      "text": "RT @user: You next #Apple #iPhone7, #iPhone8 for not even $100 $AAPL in Magic Black https://t.co/bYPPG6yHQU @user @user @user ",
      "truncated": false,
      "entities": {
        "hashtags": [
          { "text": "Apple", "indices": [26,32] },
          { "text": "iPhone7", "indices": [33,41] },
          { "text": "iPhone8", "indices": [43,51] }
        ],
        "symbols": [
          { "text": "AAPL", "indices": [70,75] }
        ],
        "user_mentions": [
          { "screen_name": "user", "name": "user", "id": id, "id_str": "id", "indices": [3,15] },
          { "screen_name": "user", "name": "user", "id": id, "id_str": "id", "indices": [116,124] },
          { "screen_name": "user", "name": "user", "id": id, "id_str": "id", "indices": [125,135] }
        ],
        "urls": [
          { "url": "https://t.co/bYPPG6yHQU", "expanded_url": "http://www.oukitel.com/products/u/u20-plus-58.html", "display_url": "oukitel.com/products/u/u20 ", "indices": [91,114] }
        ]
      },
      "metadata": {
        "iso_language_code": "en",
        "result_type": "recent"
      },
      "source": "<a href='\"http://twitter.com/download/android\"' rel='\"nofollow\">Twitter for Android</a>",
      "in_reply_to_status_id": null,
      "in_reply_to_status_id_str": null,
      "in_reply_to_user_id": null,
      "in_reply_to_user_id_str": null,
      "in_reply_to_screen_name": null,
      "user": {
        "id": 3581738234,
        "id_str": "3581738234",
        "name": "user",
        "screen_name": "user",
        "location": "Budapest, Hungary",
        "description": "Former ... from Frankfurt, #Germany - Doing God's Work for #Hungary in Hong Kong, #China",
        "url": "https://t.co/heqmfM3SUm",
        "entities": {
          "url": {
            "urls": [
              { "url": "https://t.co/heqmfM3SUm", "expanded_url": "http://bse.hu/", "display_url": "bse.hu", "indices": [0,23] }
            ]
          },
          "description": {
            "urls": []
          }
        },
        "protected": false,
        "followers_count": 855,
        "friends_count": 1318,
        "listed_count": 620,
        "created_at": "Wed Sep 16 10:36:49 +0000 2015",
        "favourites_count": 24195,
        "utc_offset": -28800,
        "time_zone": "Pacific Time (US & Canada)",

```

Listing B.1: Truncated JSON text returned by Twitter REST API

Table 26 presents ten StockTwits messages and the respective classification assigned by their authors and the SA procedure applied in this work.

This section also shows two examples of financial data utilized in this thesis. Table 27 shows a short sample (40 trading days) of SP500 data collected from Thomson Reuters Datastream. The *RI* column contains return index values and *VO* column has the turnover by volume values.

Tables 28 presents a short sample (40 trading days) of PInd returns downloaded from Prof. Kenneth French webpage http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

Appendix B. Data

Table 25.: Some tweet information converted from JSON format

Variable	Content
created_at	Wed Jan 18 13:15:30 +0000 2017
id	821707589226352640
id_str	821707589226352640
text	RT @user: You next #Apple #iPhone7, #iPhone8 for not even 100AAPL in Magic Black https://t.co/bYPPG6yHQU @user @user @user
truncated	FALSE
entities.hashtags.text	Apple
entities.hashtags.indices1	26
entities.hashtags.indices2	32
entities.hashtags.text	iPhone7
entities.hashtags.indices1	33
entities.hashtags.indices2	41
entities.hashtags.text	iPhone8
entities.hashtags.indices1	43
entities.hashtags.indices2	51
entities.symbols.text	AAPL
entities.symbols.indices1	70
entities.symbols.indices2	75
entities.user_mentions.screen_name	user
entities.user_mentions.name	user
entities.user_mentions.id	id
entities.user_mentions.id_str	id
entities.user_mentions.indices1	3
entities.user_mentions.indices2	15
entities.user_mentions.screen_name	user
entities.user_mentions.name	user
entities.user_mentions.id	id
entities.user_mentions.id_str	id
entities.user_mentions.indices1	116
entities.user_mentions.indices2	124
entities.user_mentions.screen_name	user
entities.user_mentions.name	user
entities.user_mentions.id	id
entities.user_mentions.id_str	id
entities.user_mentions.indices1	125
entities.user_mentions.indices2	135
entities.urls.url	https://t.co/bYPPG6yHQU
entities.urls.expanded_url	http://www.oukitel.com/products/u/u20-plus-58.html
entities.urls.display_url	oukitel.com/products/u/u20
entities.urls.indices1	91
entities.urls.indices2	114
metadata.iso_language_code	en
metadata.result_type	recent
source	http://twitter.com/download/android
user.id	id
user.id_str	id
user.name	user
user.screen_name	textituser
user.location	Budapest, Hungary
user.description	Former (...)
user.url	https://t.co/heqmfM3SUM
user.entities.url.urls.url	https://t.co/heqmfM3SUM
user.entities.url.urls.expanded_url	http://bse.hu/
user.entities.url.urls.display_url	bse.hu
user.entities.url.urls.indices1	0
user.entities.url.urls.indices2	23
user.protected	FALSE
user.followers_count	855
user.friends_count	1318
user.listed_count	620
user.created_at	Wed Sep 16 10:36:49 +0000 2015
user.favourites_count	24195
user.utc_offset	-28800
user.time_zone	Pacific Time (US & Canada)
user.geo_enabled	FALSE
user.verified	FALSE
user.statuses_count	39428
user.lang	en
user.contributors_enabled	FALSE
user.is_translator	FALSE

Table 26.: Short sample (10 messages) of StockTwits messages, author supplied label and SA classification

StockTwits Message	Author Label	SA Label
\$FIVE settin up well for a big breakout in the coming weeks.	Bullish	Bullish
Thats why it rallied kinda late. You might want to exit now before it retraces 10 percent tomorrow.	Bearish	Bearish
\$SHLD \$LNKD \$CMG calls on our morning watchlist before the open. Buy when its cold sell while is steaming hot.	Bullish	Bearish
Let the fat cats buy the 3 dip on ocz! lol.	Bullish	Bullish
\$AAPL will continue strong into close.	Bullish	Bullish
\$ES.F balanced day, so short 97.75, 4 ticks risk, looking for 94.50. Expecting buyers at 96, may join them there.	Bearish	Bearish
Selling half my small \$LNKD today that remained. Congrats to LinkedIn team on their focus and winning hearts and minds. Love the product.	Bullish	Bullish
\$AAPL the blackberry, samsung, lg, nor any other phone will be hotter in higher demand than \$AAPL. Who camps for other phones? No <i>one</i> .	Bullish	Bullish
\$ALNY weird action. Seems like it's artificially stuck at 25. \$\$.	Bearish	Bullish
\$SPX macro what we have here is a central planning, socialist stock market! God bless America!	Bearish	Bearish

The words and expressions included in the stock market lexicon after the pre-processing operations (e.g., lemmatization) are in **bold**. Negated segments are in *italics*

Appendix B. Data

Table 27.: Short sample (40 trading days) of SP500 data collected from Thomson Reuters Datastream

Date	RI	VO
2013-02-22	2670.36	505050
2013-02-25	2621.46	627165
2013-02-26	2637.89	569354
2013-02-27	2672.11	479813
2013-02-28	2669.92	735425
2013-03-01	2676.18	522709
2013-03-04	2688.54	490188
2013-03-05	2714.35	499572
2013-03-06	2718.07	490727
2013-03-07	2723.27	489028
2013-03-08	2735.67	479459
2013-03-11	2744.81	437760
2013-03-12	2738.35	457795
2013-03-13	2742.87	409149
2013-03-14	2758.3	494834
2013-03-15	2753.82	1426617
2013-03-18	2738.68	495349
2013-03-19	2732.09	535005
2013-03-20	2750.42	499141
2013-03-21	2727.69	505315
2013-03-22	2747.25	491219
2013-03-25	2738.08	522845
2013-03-26	2760.32	404453
2013-03-27	2758.85	407533
2013-03-28	2770.05	661078
2013-03-29	2770.05	NA
2013-04-01	2757.83	413574
2013-04-02	2772.11	462829
2013-04-03	2743.38	576828
2013-04-04	2754.67	471727
2013-04-05	2742.86	547925
2013-04-08	2761.05	448688
2013-04-09	2770.84	511134
2013-04-10	2804.81	510781
2013-04-11	2815.02	521191
2013-04-12	2807.1	529328
2013-04-15	2742.63	703753
2013-04-16	2781.86	548572
2013-04-17	2742.01	673253
2013-04-18	2723.72	615946

Table 28.: Short sample (40 trading days) of PInd returns collected from Prof. Kenneth French webpage

Date	NoDur	Durbl	Manuf	Enrgy	HiTec	Telcm	Shops	Hlth	Utils	Other
2015-01-26	0.68	0.42	1.07	2	0.86	0.56	0.78	1.97	0.18	0.6
2015-01-27	-0.36	-0.48	-0.61	1.43	-0.66	-0.62	-0.41	0.39	0.03	-0.52
2015-01-28	-1.04	-1.78	-1.21	-5.13	-1.03	-1.91	-1.27	-1.26	-1.36	-1.45
2015-01-29	1.09	2	0.8	-0.12	0.89	0.75	1.23	1.4	1.35	0.77
2015-01-30	-2.34	-1.83	-1.3	2.1	-1.51	-1.38	-1.7	-1.57	-1.91	-1.33
2015-02-02	0.82	0.98	1.27	5.49	0.76	1.14	0.48	-0.4	0.9	1.15
2015-02-03	1.65	2.04	2.49	6.69	1.7	2.5	1.8	0	0.97	1.49
2015-02-04	-0.49	0.46	-0.98	-2.06	-0.14	-0.02	-0.44	-0.9	-1.28	-0.23
2015-02-05	0.61	0.94	1.43	3.55	1.3	0.72	1.28	1.88	1.22	1.19
2015-02-06	-0.2	0.85	0.14	0.85	0.05	0.13	0.22	-0.38	-3.34	0.27
2015-02-09	-0.71	-0.19	0.02	1.91	-0.32	-0.41	-0.75	-0.22	-0.87	-0.55
2015-02-10	0.42	-0.13	0.12	-2.43	0.76	0.51	0.55	1.18	1.47	0.3
2015-02-11	0.09	-0.01	0.04	-0.71	0.04	-0.13	0.21	0.1	-1.75	-0.02
2015-02-12	0.56	0.59	1.41	2.11	1.04	0.94	0.42	1.2	0.16	0.94
2015-02-13	0.53	0.51	1.08	3.09	0.84	1.06	0.5	1.18	-0.89	0.48
2015-02-17	-0.29	0.61	0.15	1.82	0.03	-0.07	0.04	1.64	-0.22	0.11
2015-02-18	0.42	0.02	0.19	-1.42	0.02	-0.21	0.4	1.05	1.95	-0.19
2015-02-19	0.46	0.41	0.01	-0.65	0.48	0.05	0.01	0.46	-0.39	0.12
2015-02-20	0.47	-0.41	0.22	-0.8	0.15	0.58	-0.05	0.48	-0.07	0.06
2015-02-23	-0.24	0.19	-0.48	-1.48	-0.5	-0.79	-0.45	0.38	0.55	-0.27
2015-02-24	0.26	-0.2	0.75	-0.22	0.6	0.34	0.3	-0.61	0.48	0.78
2015-02-25	0.17	0.63	0.1	1.16	0.67	0.37	0.11	1.41	-0.84	0.13
2015-02-26	-0.16	0.06	0.32	-2.4	0.1	-0.08	-0.1	1.17	-0.44	0.05
2015-02-27	-0.22	-0.47	-0.51	-0.66	-0.19	-0.52	-0.39	-0.87	-0.34	-0.43
2015-03-02	0.6	0.84	0.85	-1.78	1.05	1.16	0.59	1.18	-1.41	0.65
2015-03-03	-0.16	-0.66	-0.73	0.32	-0.74	-0.25	-0.38	-0.43	0.38	-0.44
2015-03-04	-0.34	-0.77	-0.65	-0.21	-0.29	-0.78	-0.84	0.93	-0.68	-0.49
2015-03-05	-0.21	0.67	0.16	-0.88	0.3	0.72	0.22	1.11	0.5	0.16
2015-03-06	-1.34	-1.09	-1.19	-1.88	-1.16	-0.59	-0.84	-0.61	-2.37	-0.52
2015-03-09	-0.07	0.27	0.35	-2.31	-0.12	0.4	0.37	-0.16	0.25	0.3
2015-03-10	-1.29	-1.25	-1.72	-3.57	-1.51	-2.05	-1.23	-0.63	-0.55	-1.18
2015-03-11	-0.32	-0.16	0.36	0.12	0.15	0.04	0.25	0.27	-0.3	0.51
2015-03-12	1.8	1.15	1.24	-0.88	1.19	2.18	1.72	1.08	1.78	1.64
2015-03-13	-0.62	-0.74	-0.44	-1.47	-0.37	-0.41	-0.57	-0.21	-0.95	-0.33
2015-03-16	0.1	-0.1	-0.12	-0.76	0.3	0.2	0.31	0.61	1.11	0.23
2015-03-17	0.18	0.17	-0.08	0.06	0.11	0.55	0.16	0.49	0.06	0.12
2015-03-18	1.01	1.25	1.18	3.88	0.86	1.25	0.58	0.26	2.43	0.15
2015-03-19	0.51	0.1	-0.59	-2.49	0.43	-0.6	0.35	1.84	-0.9	-0.16
2015-03-20	0.72	0.67	0.65	0.84	0.29	0.53	0.71	-1.01	1.17	0.81
2015-03-23	1.14	0.44	0.53	0.04	0.18	0.53	0.37	-0.62	0	0.18