Universidade do Minho
Escola de Engenharia

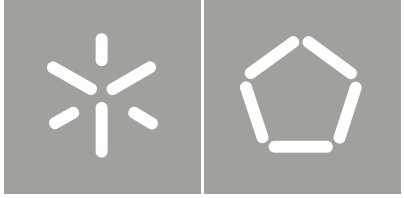
Nuno Miguel da Rocha Oliveira

Mining Microblogging Data to Model and Forecast Stock Market Behavior

October 2013

Nuno Miguel da Rocha Oliveira Mining Microblogging Data to Model and Forecast Stock Market Behavior

UMinho | 2013



Universidade do Minho
Escola de Engenharia

Nuno Miguel da Rocha Oliveira

Mining Microblogging Data to Model and Forecast Stock Market Behavior

Master Thesis
Master in Information Systems Engineering and Management

Work performed under the guidance of Professor:
Paulo Alexandre Ribeiro Cortez

October 2013

DECLARAÇÃO

Nome: Nuno Miguel da Rocha Oliveira

Endereço electrónico: nunomroliveira@gmail.com

Telefone: 936607860

Número do Bilhete de Identidade: 11037993

Título tese: Mining Microblogging Data to Model and Forecast Stock Market Behavior

Orientador: Paulo Alexandre Ribeiro Cortez

Ano de conclusão: 2013

Designação do Mestrado: Mestrado em Engenharia e Gestão de Sistemas de Informação

Universidade do Minho, ___/___/_____

Assinatura: _____

Acknowledgments

To my supervisor, Paulo Cortez for the invaluable guidance, support, collaboration and motivation.

I wish also to thank Nelson Areal, from the Department of Management and Economics of University of Minho, for his collaboration with this work.

Mining Microblogging Data to Model and Forecast Stock Market Behavior

Abstract

The analysis of microblogging data may disclose relevant signals of investor sentiment and attention that can be useful to model and predict stock market variables (Bollen, Mao, & Zeng, 2011; Mao, Counts, & Bollen, 2011; Oh & Sheng, 2011; Sprenger & Welpe, 2010). Moreover, microblogging data can provide sentiment and attention indicators in a more rapid and cost-effective manner than traditional sources (e.g., large scale surveys).

In this project, we assessed the information content of microblogging data for explaining stock market variables. We created several indicators using Twitter data from nine major technological companies and analyzed their value when modeling returns, trading volume and volatility. Sentiment indicators were produced by exploring 5 popular lexical resources and two novel lexicons (emoticon based and the merge of all 6 lexicons) while attention indicators were based on the posting volume.

Despite the short period analyzed (32 days), interesting results were obtained when measuring the value of using posting volume for fitting trading volume and volatility. However, we found scarce evidence that sentiment indicators can explain stock returns.

Análise de Dados de Microblogs para Modelar e Prever o Comportamento do Mercado de Ações

Resumo

A análise de dados de microblogging pode revelar sinais relevantes do sentimento e atenção do investidor que podem ser úteis para modelar e prever variáveis do mercado de ações (Bollen et al., 2011; Mao et al., 2011; Oh & Sheng, 2011; Sprenger & Welpe, 2010). Adicionalmente, esta fonte de dados pode fornecer indicadores de sentimento e atenção de uma forma mais rápida e económica que fontes tradicionais (e.g., sondagens).

Neste projecto, avaliamos o conteudo informativo dos dados de microblogging para explicar variáveis de mercados de ações. Criamos vários indicadores utilizando dados do Twitter sobre nove grandes empresas tecnológicas e analisamos o seu valor para modelar rendibilidade, volume de transação e volatilidade. Os indicadores de sentimento foram produzidos utilizando cinco recursos léxicos populares e dois novos lexicons (emoticons e união dos seis lexicons) enquanto que os indicadores de atenção se basearam no número de tweets.

Apesar do curto período de tempo analisado (32 dias), obtivemos resultados interessantes na utilização do número de tweets para modelar o volume de transação e volatilidade. Contudo, encontramos evidência escassa que os indicadores de sentimento podem explicar as rendibilidades das ações.

Table of contents

ACKNOWLEDGMENTS ..... **III**

ABSTRACT ..... **V**

RESUMO ..... **VII**

ACRONYMS/ NOTATION ..... **XI**

LIST OF FIGURES ..... **XIII**

LIST OF TABLES ..... **XV**

1. INTRODUCTION **1**

1.1. MOTIVATION 1
1.2. OBJECTIVES 2
1.3. ORGANIZATION 3

2. LITERATURE REVIEW **5**

2.1. INTRODUCTION 5
2.2. BUSINESS INTELLIGENCE 5
2.3. DATA MINING (DM) 7
2.3.1. Neural Networks (NN) *8*
2.3.2. Support Vector Machines (SVM) *8*
2.3.3. Decision Trees *9*
2.3.4. Naïve Bayes *10*
2.4. TEXT MINING (TM) 11
2.4.1. Traditional TM Framework *12*
2.4.1.1. Text Preprocessing 12
2.4.1.2. Text Representation 12
2.4.1.3. Knowledge Discovery 13
2.4.2. Information Extraction (IE) *13*
2.4.3. Text Classification (TC) *14*
2.4.4. Information Summarization *15*
2.4.5. Text Clustering *17*
2.4.6. TM in Social Media *17*
2.4.6.1. Time Sensitivity 18
2.4.6.2. Short Length 19
2.4.6.3. Unstructured Phrases 19
2.4.6.4. Abundant Information 19
2.4.7. NLP Resources *20*
2.4.7.1. Part-of-Speech (POS) Tagger 20
2.4.7.2. Constituency Parser 20
2.4.7.3. Dependency Parser 21
2.4.7.4. Shallow Parser 21
2.4.7.5. Opinion Lexicon 21
2.5. OPINION MINING (OM) 22
2.5.1. Opinion Definition *23*
2.5.2. Document-level Sentiment Classification *24*
2.5.3. Sentence-level Sentiment Classification *24*
2.5.4. Aspect-level Sentiment Classification *25*
2.5.5. Mining Comparative Opinions *25*

2.6. UTILIZATION OF MICROBLOGGING DATA TO MODEL AND FORECAST STOCK MARKET VARIABLES 26
2.6.1. Investor Sentiment and Attention *27*
2.6.2. Microblogging data *27*
2.6.3. Stock Market Variables *29*
2.6.4. Related work using microblogging data for stock market prediction *29*
2.6.5. Related work applying other sources of web social data *31*
2.6.5.1. Internet Searches 34
2.6.5.2. Blogs 37
2.6.5.3. Message Boards 38
2.6.6. Summary *40*

3. EXPERIMENTS ON MODELING STOCK MARKET BEHAVIOR USING INVESTOR SENTIMENT ANALYSIS AND POSTING VOLUME FROM TWITTER 43

3.1. INTRODUCTION 43
3.2. MATERIALS AND METHODS 44
3.2.1. Twitter Data *44*
3.2.2. Stock Market Data *47*
3.2.3. Sentiment Analysis Methods *47*
3.2.3.1. Pre-processing 47
3.2.3.2. Lexical Resources 48
3.2.3.3. Sentiment Analysis Approaches 49
3.2.4. Regression Models *50*
3.2.4.1. Returns 51
3.2.4.2. Trading Volume 51
3.2.4.3. Volatility 52
3.2.5. Evaluation *52*
3.3. RESULTS 53
3.3.1. Returns *53*
3.3.2. Volatility *55*
3.3.3. Trading Volume *58*
3.4. DISCUSSION 62

4. CONCLUSIONS 65

4.1. SUMMARY 65
4.2. DISCUSSION 66
4.3. FUTURE WORK 67

REFERENCES 69

Acronyms/ Notation

AMZN	Amazon
API	Application Programming Interface
BI	Business Intelligence
BOW	Bag Of Words
DM	Data Mining
GI	General Inquirer
GOOG	Google
HTML	HyperText Markup Language
IDF	Inverse Document Frequency
IE	Information Extraction
INTC	Intel
MAPE	Mean Absolute Percentage Error
ML	Machine Learning
MPQA	Multi-Perspective Question Answering
MSFT	Microsoft
MSOL	Macquarie Semantic Orientation Lexicon
NER	Named Entity Recognition
NLP	Natural Language Processing
NN	Neural Networks
OL	Opinion Lexicon
OM	Opinion Mining
POS	Part-of-Speech
RAE	Relative Absolute Error
RE	Relation Extraction
REST	Representational State Transfer
S&P	Standard & Poor's
S1	First Sentiment Analysis Approach
S2	Second Sentiment Analysis Approach
SVM	Support Vector Machine
SWN	SentiWordNet
TC	Text Classification

TF	Term Frequency
TF-IDF	Term Frequency - Inverse Document Frequency
TM	Text Mining
WIMS	International Conference on Web Intelligence, Mining and Semantics

List of Figures

FIGURE 1. MAXIMUM MARGIN HYPERPLANE AND SUPPORT VECTORS 9

FIGURE 2. LITERATURE MAP ABOUT MINING MICROBLOGGING DATA TO MODEL AND FORECAST STOCK MARKET BEHAVIOR 32

FIGURE 3. LITERATURE MAP ABOUT MINING WEB DATA TO MODEL AND FORECAST STOCK MARKET BEHAVIOR 35

FIGURE 4. SCHEMATIC OF THE ADOPTED EXPERIMENTATION SETUP ... 45

FIGURE 5. TOTAL NUMBER OF TWEETS COLLECTED FOR THE NINE SELECTED TECHNOLOGICAL COMPANIES 46

FIGURE 6. INTC RETURNS AND PREDICTIVE VALUES 56

FIGURE 7. VOLATILITY AND FITTED VALUES FOR AMZN 59

FIGURE 8. TRADING VOLUMES AND FITTED VALUES FOR AMD 61

List of Tables

TABLE 1. BI APPLICATIONS 6

TABLE 2. INTERNET TRAFFIC REPORT BY ALEXA ON SEPTEMBER 30TH, 2013 18

TABLE 3. RESEARCH ABOUT MINING MICROBLOGGING DATA TO MODEL AND FORECAST STOCK MARKET BEHAVIOR 33

TABLE 4. RETURNS USING S1 FEATURES RESULTS 54

TABLE 5. RETURNS USING S2 FEATURES RESULTS 54

TABLE 6. VOLATILITY $\boldsymbol{R2}$ RESULTS 57

TABLE 7. VOLATILITY RAE RESULTS 57

TABLE 8. VOLUME $\boldsymbol{R2}$ RESULTS 60

TABLE 9. VOLUME RAE RESULTS 60

1. Introduction

1.1. Motivation

The analysis and prediction of stock market behavior is a focus of researchers' attention for a long time. A better prediction of variables related to behavioral aspects of the stock market can lead to valuable benefits, such as the formulation of better strategies for investment portfolios.

The Efficient Market Hypothesis defends that investors act as rational agents and all existing information is reflected immediately in stock prices. However, financial research has shown that financial decisions are significantly driven by emotion and mood (Nofsinger, 2005) and investors' attention can have an effect on asset prices and dynamics (Hirshleifer & Teoh, 2003; Merton, 1987).

Sentiment and attention indicators created from microblogging data may potentially improve the prediction of stock market variables. The community of users that utilizes these microblogging services to share information about stock market issues has grown and is potentially more representative of all investors. The analysis of its contents can allow the extraction of important signals of sentiment from investors regarding several stock market issues. Moreover, microblogging data is readily available at low cost permitting a faster and less expensive creation of indicators, compared to traditional sources (e.g., large-scale surveys), and can also contain new information that is not present in historical quantitative financial data. Furthermore, the small size of the message (maximum 140 characters) and the usage of cashtags (a hashtag identifier for financial stocks) can make it a less noisy source of data. Finally, users post very frequently, reacting to events in real-time and allowing a real-time assessment that can be exploited during the trading day.

Mining microblogging data to model and forecast stock market behavior is a very recent research topic that has presented promising results (Bollen et al., 2011; Fuehres, Zhang, & Gloor, 2011; Mao et al., 2011; Oh & Sheng, 2011; Ruiz, Hristidis, Castillo, Gionis, & Alejandro, 2012; Sprenger & Welpe, 2010). In such literature, it is argued that a model that accounts for investor sentiment and attention can provide a better explanation of stock market behavior and potentially be used to predict key stock market variables, such as returns, volatility and trading volume.

1.2. Objectives

This research topic is very recent and the research results are not consolidated. In this project, we intend to extend this investigation by exploring different sentiment analysis methods and concentrating in a specific sector (i.e. technological sector). The main research objectives are:

- Perform a rigorous analysis of the state of the art related with mining microblogging data to model and forecast stock market behavior. This assessment should contribute to verify the relationship between microblogging features and stock market variables and also to identify opportunities to improve results in this research topic.

- Evaluate the relevance of a sentiment analysis method, unexplored in this topic, to create indicators of investor sentiment. We will produce sentiment indicators using five popular and large lexical resources and two new proposed lexicons: emoticons; and ALL, which merges the six remaining resources. If this method is more effective, sentiment indicators will have added value to stock market models.

- Assess the information content of Twitter data for explaining some stock market variables in the technological sector. Research in this

area has studied indexes or stocks from a wide range of sectors. We will consolidate existing research by focusing on individual stocks of a specific sector that has a substantial posting volume. Therefore, Twitter data may be more representative of investors sentiment and attention regarding these stocks and have more informative content for the modeling of stock market behavior.

1.3. Organization

This document is divided into four chapters:

- the first introduces the theme, presents the motivation for this project and enumerates the main research objectives;
- the second chapter presents the literature review that provided the theoretical support for the implementation of the work plan;
- the third chapter describes the whole research project, namely the data and methods applied and the obtained results;
- the final chapter summarizes the research project, presents the main conclusions and recommends future work.

2. Literature Review

2.1. Introduction

Mining microblogging data to model and forecast stock market behavior requires a diverse set of knowledge and skills from various fields of study. Therefore, it is convenient to perform a thorough study of diverse disciplines. This body of knowledge provides the theoretical background for the execution of the project and permits the identification of research opportunities to improve results.

In this chapter, we provide an overview of diverse fields of study related to this topic and describe the state of the art about mining microblogging data to forecast stock market behavior. We chose to highlight Text Mining and Opinion Mining due to their relevance in this research topic.

2.2. Business Intelligence

Business intelligence (BI) is an umbrella term that includes architectures, tools, databases, applications, and methodologies. BI major objective is to enable interactive access (sometimes in real time) to data, enable manipulation of these data, and to provide business managers and analysts the ability to conduct appropriate analysis (Turban, Sharda, Aronson, & King, 2007). It seeks to satisfy managers' need of the right information at the right time, in the right place. The process of BI is based on the transformation of data to information, then to decisions, and finally to actions. Some BI applications are presented in Table 1.

Table 1. BI Applications

Analytic Application	Business Question	Business Value
Customer segmentation	What market segments do my customers fall into, and what are their characteristics?	Personalize customer relationships for higher customer satisfaction and retention.
Propensity to buy	Which customers are most likely to respond to my promotion?	Target customers based on their need to increase their loyalty to your product line. Also, increase campaign profitability by focusing on the most likely to buy.
Customer profitability	What is the lifetime profitability of my customer?	Make individual business interaction decisions based on the overall profitability of customers.
Fraud detection	How can I tell which transactions are likely to be fraudulent?	Quickly determine fraud and take immediate action to minimize cost.
Customer attrition	Which customer is at risk of leaving?	Prevent loss of high-value customers and let go of lower-value customers.
Channel optimization	What is the best channel to reach my customer in each segment?	Interact with customers based on their preference and your need to manage cost.

Source: (Zaima & Kashner, 2003)

A BI system may include several components, such as: a **Data Warehouse** with its source data; business analytics that has a collection of tools for manipulating and analyzing the data in the Data Warehouse, including Data Mining; **Business Performance Management** for the monitoring and analysis of performance; and a user interface (such as the dashboard) (Turban et al., 2007).

In this review we focus on the analytical component, namely Data Mining and its Text Mining variant, that are applied to mine sentiment and attention indicators from social media data to model and forecast stock market behavior.

2.3. Data Mining (DM)

DM is a process that uses database, statistical, mathematical, artificial intelligence, and Machine Learning (ML) techniques to extract and identify useful information and subsequent knowledge from raw data (Fayyad, Piatetsky-shapiro, & Smyth, 1996). It finds mathematical patterns that can be rules, affinities, correlations, trends, or prediction models. DM offers organizations an indispensable decision-enhancing environment to exploit new opportunities by transforming data into a strategic weapon (Turban et al., 2007). Here we present some DM common goals (Fayyad et al., 1996):

Classification: The objective is to assign the correct label to unclassified records using a model trained with a pre-classified data set.

Clustering: It aims to divide a database into segments whose members share similar qualities. Unlike classification, the clusters are unknown when the algorithm starts.

Association: It seeks to establish relationships about items that occur together in a given record. One of the main applications of this technique is the analysis of sales transactions.

Regression: The objective is to map the attributes that characterize an item into a target continuous value. There are linear and nonlinear regression techniques.

Sequence Discovery: The goal is to identify associations over time. Thus, it can contribute to understand behavior over time and have several applications such as marketing or fraud detection.

Visualization: The objective is to enable an easily understandable presentation of data. It converts complex data characteristics on clear patterns to allow users a better visualization of the complicated discoveries made in the process of DM.

The most widely used DM techniques in mining microblogging data to forecast stock market behavior are described in the next subsections.

2.3.1. Neural Networks (NN)

NN are inspired in the human nervous system. The model contains a system of interrelated and parallel computational units called neurons, organized in layers, forming a network. Most NN are composed by three types of layers: input, hidden, and output. The value of each neuron is calculated by linearly combining the value from neurons of the preceding layer and by applying an activation function. The activation function is selected considering the nature of the data and the assumed distribution of target variables (Bishop, 2006).

NN have disadvantages as well as advantages. They tend to be most effective where there are a very large number of variables and the relationships between them are complex and imprecise. A NN can easily be implemented in a parallel environment, with each node doing its calculations on a different processor. However, it is usually very difficult to provide a good rationale for the predictions made by a NN. Additionally, NN require more computational resources (e.g. training effort) when compared with simpler modeling techniques. In particular, the time needed for training tends to increase as the volume of the data increases, and in general, NN cannot be trained on very large databases (Turban et al., 2007).

2.3.2. Support Vector Machines (SVM)

SVM is a popular ML method for tasks such as classification and regression. SVM use a linear model to implement nonlinear class boundaries by mapping input vectors nonlinearly into a high-dimensional feature space (Smola & Schölkopf, 2004). They are based on an algorithm that finds a special kind of linear model: the maximum margin hyperplane. This hyperplane is the one that gives the greatest separation between datasets classes that are linearly separable, permitting a more

accurate classification of the training instances. An example is shown in Figure 1.

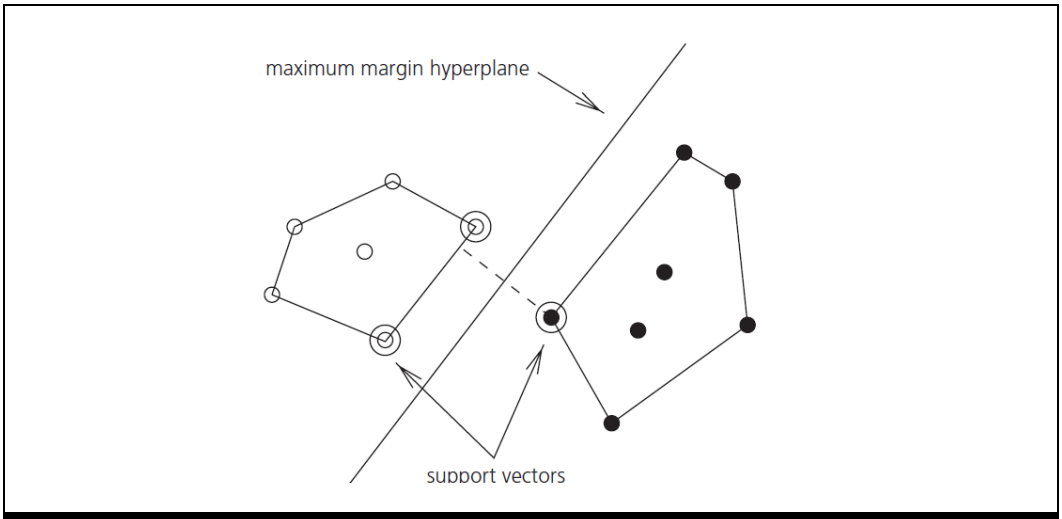


Figure 1. Maximum margin hyperplane and support vectors
Source: (Witten & Frank, 2005)

Support vectors are the instances that are closest to the maximum margin hyperplane. They permit an easy construction of the mentioned hyperplane. All other training instances become irrelevant (Witten & Frank, 2005).

After training, new examples are mapped to the same space and, depending on which side of the hyperplane they fall, their category is predicted. Several other SVM variants have been proposed, such as Support Vector Regression, which can be applied to regression tasks.

2.3.3. Decision Trees

Decision trees are comprised of an hierarchy of if-then statements and are used in classification and clustering methods (Turban et al., 2007). They can be defined as a root followed by internal nodes, ending in leaf nodes that represent the final class choice for a pattern. Each node is labeled with a question that represents a test on an attribute and each branch represents a response to that question. The questions should be the ones that best divide the training records. A new instance is classified by

answering these successive questions and following the corresponding branches. The assigned class corresponds to the leaf node.

Decision trees break down problems into increasingly discrete subsets by working from generalizations to increasingly more specific information. The very general algorithm for building a decision tree is as follows (Turban et al., 2007):

1. Create a root node and select a splitting attribute.
2. Add a branch to the root node for each split candidate value and label.
3. Take the following iterative steps:
    a. Classify data by applying the split value.
    b. If a stopping point is reached, then create a leaf node and label it. Otherwise, build another subtree.

There are several algorithms for creating decision trees, like ID3, C4.5, C5 from ML, Classification and Regression Trees from statistics, and Chi-squared Automatic Interaction Detector from pattern recognition. Algorithms differ primarily in terms of the choice of splitting attributes, the order of splitting attributes, the number of splits, the tree structure, the stopping criteria, and the pruning of the tree.

2.3.4. Naïve Bayes

A Naïve Bayes classifier is a probabilistic approach based on the Bayes' rule with strong independence assumptions. The Bayes' formula is:

$$P(c_k|\mathbf{x}) = P(c_k) \times \frac{P(\mathbf{x}|c_k)}{P(\mathbf{x})} \quad (1)$$

where c_k is a specific class (e.g., theme) and **x** is a vector of feature values representing an event (e.g., text document). In short, this rule states that the probability of a event belonging to a particular class depends on the conditional probability of its features occurring in a class (Lewis, N’{e}dellec, & Rouveirol, 1998). Bayes' rule suggests that the

estimation of $P(c_k|\mathbf{x})$ can be achieved by calculating $P(\mathbf{x}|c_k)$, $P(c_k)$ and $P(\mathbf{x})$.

Bayes classifiers use a set of labeled training instances to calculate Bayes-optimal estimates of the model parameters. Then, new examples are classified with the highest valued class according to the generative model and the Bayes' rule.

Naïve Bayes methods simplify the procedure by assuming that all attributes are independent of each other given the context of the class. Thus, the parameters for each attribute can be learned separately, facilitating the learning process. This assumption is usually false in real-world situations, however Naïve Bayes often obtains very good results (McCallum & Nigam, 1998).

2.4. Text Mining (TM)

TM intends to discover important information from unstructured or less structured text files. The main problem is that the information is not couched in a manner that is amenable to automatic processing. TM strives to bring it out in a form suitable for consumption by computers (Witten & Frank, 2005). It usually involves the process of structuring text and then extract patterns and trends from the structured data. TM is an interdisciplinary field, applying techniques from diverse areas such as Natural Language Processing (NLP), DM, ML or Information Retrieval (Aggarwal & Zhai, 2012a).

The explosion of textual contents created in social networks and web has increased the need for algorithms able to discover interesting knowledge from the data in a dynamic and scalable way (Aggarwal & Zhai, 2012a). TM can support users to quickly analyze information and make better decisions. It is applied in a wide range of domains such as business, security, marketing, research or biomedicine.

In the following subsections, we will describe a generic TM framework, some of the most relevant TM applications and common NLP resources.

2.4.1. Traditional TM Framework

TM processes need to perform diverse operations in order to transform text documents from an raw and unstructured format into a structured representation, and, then, to discover useful knowledge. Despite the variety of possible TM processes, an usual framework can be generically characterized by three main phases.

2.4.1.1. Text Preprocessing

This task aims to prepare and facilitate the next TM phases, transforming the input documents without losing important information. Some frequent problems of TM systems, such as the high dimensionality and sparsity of the features, are addressed in this phase (Aggarwal & Zhai, 2012a). Traditional text preprocessing methods are:

- Stop word removal: elimination of common words that are considered meaningless (e.g. the, a);
- Stemming (Porter, 1980): word replacement by their stem, base or root form. Many words can be represented by the same feature (e.g. the stem of "process", "processing" or "processed" is "process").

The selection of the preprocessing methods depend on the succeeding TM steps. For example, applications that require an appropriate syntactical analysis should discard a stop word removal.

2.4.1.2. Text Representation

The objective of this phase is to identify a set of features that can represent the whole content. These representational models can be constituted by diverse type of features such as characters, words, syntactic tags or concepts (Feldman & Sanger, 2007). A major challenge is to obtain a group of features that can, simultaneously, contain the

appropriate semantic information and be computationally efficient for the subsequent knowledge discovery. Various techniques from information extraction and computational linguistics can be adapted and applied in this phase. External knowledge sources such as dictionaries, ontologies or knowledge bases can also be utilized to generate features semantically richer (Feldman & Sanger, 2007).

The most common representation is the "Bag of words" (BOW). This approach transforms text documents into sparse numeric vectors constituted by numerical values representing the frequency of each word in the document (e.g. TF-IDF). It is a simple but limited method that ignores the linguistic structure within the text, preventing more rigorous and meaningful analysis and mining (Feldman & Sanger, 2007).

2.4.1.3. Knowledge Discovery

The final task receives the representational models from the previous phase and seeks to discover important knowledge. TM can rely on diverse existing DM methods to identify relevant patterns, connections and trends in the entire corpus (Aggarwal & Zhai, 2012a; Feldman & Sanger, 2007).

These methods may deliver a very large number of results. Thus, an important operation is to limit this overabundance by defining measures of interest. Background knowledge sources such as lexicons or knowledge bases may also be utilized to create meaningful constraints in knowledge discovery operations (Feldman & Sanger, 2007).

2.4.2. Information Extraction (IE)

IE is the task of finding structured information from text such as entities, relations or events (Feldman & Sanger, 2007; J. Jiang, 2012). It is one of the main applications of TM. Many TM algorithms use IE as a starting point because it can disclose significant semantic information and support inferences about knowledge discovered in text (J. Jiang, 2012). Two essential tasks of IE are named entity recognition (NER) and relation extraction (RE).

The objective of NER is to identify entities from text and then to classify them into a set of types (J. Jiang, 2012). It has to recognize sequences of words that correspond to real world entities (e.g., "Barack Obama", "Google", "Great Britain") and associate it to specific entity types such as person, organization or location. Other IE tasks, such as RE or event extraction, use NER as a pre-processing step (Feldman & Sanger, 2007; J. Jiang, 2012). Thus, NER is a fundamental task in IE.

RE is the task of finding and categorizing the relations between entities (J. Jiang, 2012). For example, given the sentence "In 1928, Alexander Fleming discovered penicillin.", we can extract the following relations:

- DiscovererOf(Alexander Fleming, penicillin),
- DiscoveredIn(penicillin, 1928).

IE has applications in various domains, such as:

- Biomedicine: automatically identify and classify mentions of biomedical entities from literature;
- Finance: extract detailed information about financial issues (e.g., takeovers) from news articles.
- Intelligence: discover important information related to terrorism such as people involved, the weapons used and the targets of the attacks.

2.4.3. Text Classification (TC)

The objective of TC is to classify each data instance (e.g. news, tweets, document) into a set of categories (e.g. subject, topic, sentiment polarity) (Aggarwal & Zhai, 2012b; Feldman & Sanger, 2007). Fully automated systems explicitly assign a label to each instance, whereas semi-automated versions provide a ranking of categories but the final decision is made by the user (Feldman & Sanger, 2007).

There are two main approaches to TC. The first is the knowledge engineering approach that directly encodes expert's knowledge into the

system. The second approach applies ML methods to inductively build a classifier from a set of classified data. Despite the first approach outperform ML systems in some domains (e.g. document management), the ML approach is more utilized because is less labor intensive. The knowledge engineering systems require huge amounts of highly skilled labor while ML systems only demand a set of manually classified training instances (Feldman & Sanger, 2007).

Almost all popular classification techniques have been adapted to the case of text data such as decision trees, rules, Bayes methods, nearest neighbor classifiers, SVM classifiers, and NN (Aggarwal & Zhai, 2012b).

Feature selection is an essential step for TC. Some features (e.g. words, n-grams) are more correlated to some classes than others. Thus, it is crucial for the classification process to determine the most relevant features for each class. Measures such as Gini Index, Information Gain, Mutual Information, are applied to assess the correlation between terms and categories (Aggarwal & Zhai, 2012b).

TC has a large range of applications, such as:

- news organization: automated news categorization in web portals.
- spam filtering: determine whether each e-mail is spam or not.
- sentiment classification: assign a sentiment polarity (e.g., positive, negative, neutral) to a message.
- target marketing: classify users (e.g., socioeconomic class, geographical localization) in order to implement appropriate marketing operations (e.g., ads placement).

2.4.4. Information Summarization

A frequent TM function is to provide automatic summaries of documents (Turban et al., 2007). The objective of summarization systems is to create a concise summary of the input documents that delivers the key information (Nenkova & McKeown, 2012).

There are two main approaches for text summarization. Extractive methods provide a summary composed by information units extracted from the original text. Abstractive methods apply NLP techniques to produce a summary that may contain words that are not included in the input document.

The majority of summarization systems are extractive. These methods receive a single document or a set of related documents, identify the most important sentences and aggregate them to create a summary. Usually, this process is composed by three main phases: intermediate representation, sentence scoring and summary selection (Nenkova & McKeown, 2012).

The objective of the intermediate representation is to facilitate the next summarization steps. These representations can capture the topics discussed (topic representation) or include a list of indicators (indicator representation) such as sentence length, location in the document, presence of certain phrases.

The sentence scoring phase uses the intermediate representation to assign a score indicating the importance of each sentence. In topic representation approaches, these scores intend to grade the information content that each sentence convey regarding the document topics. In indicator representation approaches, the score consider the evidence from the different indicators. Context information (e.g. document genre, web page links) can also be used to determine the importance of sentences.

Finally, the system has to select the best set of sentences to form a concise and informative summary. The combination of sentences can be selected considering variables such as sentences score or similarity with other chosen sentences in order to maximize summary importance and minimize redundancy.

2.4.5. Text Clustering

The objective of Text Clustering is to automatically find groups of similar objects in text data (Aggarwal & Zhai, 2012c). TC applications are provided with a set of pre-classified training examples while Text Clustering is an unsupervised process that has to group an unlabeled data set into meaningful clusters without any prior information (Feldman & Sanger, 2007).

Text data can have diverse representations that require different clustering algorithms. However, the TF-IDF representation is commonly used for text processing. In these representation, the term frequency (TF) for each word is normalized by the inverse document frequency (IDF). This measure reduces the importance of common terms in the collection and increases the influence of more discriminative words.

Text-specific algorithms are necessary to improve the frequent sparse and high dimensional text representation. Many information retrieval techniques can be use for this purpose.

Text clustering can be applied to several tasks, such as:

- Document Organization and Browsing: hierarchical organization of documents to facilitate the systematic browsing of the document collection.
- Corpus Summarization: creation of a coherent summary of the corpus in the form of cluster-digests (Schütze & Silverstein, 1997) or word-clusters (Baker & McCallum, 1998).
- Document Classification: utilization of clustering techniques to improve the classification accuracy of supervised applications.
- Customer segmentation: group customers with similar features.

2.4.6. TM in Social Media

Social media services (e.g., blogs, microblogs, forums) are an abundant source of opinionated text data. Nowadays, there are 5 social media

service among the top 10 sites according to statistics from Alexa[1], as shown in Table 2 (social media services are in bold).

Table 2. Internet Traffic Report by Alexa on September 30th, 2013	
Rank	Website
1	Google
2	**Facebook**
3	**Youtube**
4	Yahoo!
5	Baidu
6	**Wikipedia**
7	QQ.com
8	**LinkedIn**
9	Windows Live
10	**Twitter**

TM is a valuable tool to process the large amounts of text data produced by these social media services. It can satisfy more efficiently the information needs for various types of applications (e.g. marketing).

However, social media contents are produced in a quite different context from those produced in traditional media. Their users create contents in an environment of constant interaction and collaboration, differing from the unidirectional paradigm that exists in traditional media.

Thus, it has distinct characteristics that poses new challenges and opportunities. These features are (X. Hu & Liu, 2012):

2.4.6.1. Time Sensitivity

Social media contents are created very frequently. Some users may even post several times during a day (e.g., Facebook, Twitter) reacting to other comments about recent events. Thus, posting time is important to perform a proper contextualized analysis. For example, it is likely that messages written about a football team contain radically different opinions when written immediately after a defeat or a victory. These large number of real-

[1] www.alexa.com

time postings hold important information that can be explored for detection and monitoring of events.

2.4.6.2. Short Length

Some social media services restrict the message length (e.g. Twitter). The conciseness has advantages but also some challenges. These contents require a greater objectivity from the author and its succinctness permits more effective textual analysis than in longer standard documents. However, these messages may not provide sufficient context information.

2.4.6.3. Unstructured Phrases

Many social media messages have much less quality than traditional media contents. Users are less rigorous in the creation of text messages, producing several grammatical errors such as orthographical mistakes or incorrect punctuation. Abbreviations and acronyms are employed frequently. Some social media services (e.g., Twitter) have also specific terminology (e.g., hashtags) and structure (e.g., retweets).

These facts reduce the performance of standard NLP resources (e.g. POS and dependency taggers) and, consequently, hamper the accurate identification of the semantic meaning of these messages.

2.4.6.4. Abundant Information

In addition to the text content, social media usually contain a rich variety of information sources. These services permit to associate keywords to each message (i.e. tag information), to include connections to other users or contents (i.e. link information) and may contain other types of information such as geographical location or an author profile.

TM can extract useful information from these external sources. The analysis of the text content can be complemented by semantic clues derived from tags, links, etc. However, the abundant additional information introduces even more difficulty to an effective text collection and processing.

2.4.7. NLP Resources

NLP is a research area that develops tools and techniques that allow computers to understand, manipulate and generate natural language text in order to perform useful tasks (Jurafsky & Martin, 2000). TM algorithms constantly apply NLP resources to extract a more complete meaning representation from text. Some of most frequently applied NLP resources are described next.

2.4.7.1. Part-of-Speech (POS) Tagger

POS are linguistic categories assigned to words based on the role they play in the sentence. For example, in the sentence "Peter is fast", the POS tags could be proper noun ("Peter"), verb ("is") and adjective ("fast"). These tags provide semantic information about each word that can be useful to other TM tasks (e.g., entities are frequently nouns, opinion words are regularly adjectives).

The most common tag set is composed by seven tags: Article, Noun, Verb, Adjective, Preposition, Number, and Proper Noun. However, some POS taggers contain a much longer group of tags (e.g., Stanford POS Parser (Toutanova, Klein, Manning, & Singer, 2003) has 48 tags).

POS taggers already present excellent results in formal text documents. The state of the art Stanford POS Tagger achieves an 97.24% accuracy on the Penn Treebank WSJ corpus (Toutanova et al., 2003). However, the accuracy may drop substantially when processing more informal text documents such as those produced in social media services.

2.4.7.2. Constituency Parser

Constituency parsers perform a full syntactical analysis of sentences according to a constituency grammar. This resource identifies sequences of syntactically grouped elements (i.e. constituents) such as noun phrases, verb phrases, prepositional phrases, adjective phrases, and clauses (Feldman & Sanger, 2007). Each sentence can contain several constituents and each constituent can be composed by many words or

other phrases. Grammatical functions are also assigned to each phrase (e.g., a noun phrase may be classified as subject, object or complement).

2.4.7.3. Dependency Parser

Another type of full syntactical parsing is the dependency parsing. These parsers identify direct binary asymmetric grammatical relations (i.e. dependencies) between one word (dependent) and another word (head). The head dominates the relation, being more influential in defining the behavior of the pair. The dependent is usually the modifier, object, or complement of the head. For example, in the sentence "John writes articles", there are two dependencies: subject ("John" is the subject of the verb "writes") and object ("articles" is the object of the verb "writes"). Dependency relations are semantically richer than constituency classifications because they usually contain more grammatical roles and they already identify the head and the dependent (Covington, 2001).

2.4.7.4. Shallow Parser

Shallow parsers execute a faster but less deep analysis than full syntactical parsers. They only identify sentence components (e.g. phrases and dependencies) that are clear and unambiguous, leaving the other ones unresolved. Shallow parsers are effective for many TM applications because their level of analysis is sufficient, the results are robust and the computational effort is much smaller (Feldman & Sanger, 2007).

2.4.7.5. Opinion Lexicon

An opinion lexicon is an important resource that is employed in many sentiment classification tasks. It is composed by opinion words and phrases and the respective sentiment label (e.g. positive, negative). Their presence in the text permits to discern the sentiment orientation. For example, the sentence "The computer is good" can be easily classified as positive if the opinion lexicon contains the word "good" as positive.

There is a set of existing opinion lexicons (Baccianella, Esuli, & Sebastiani, 2010; Stone, Dunphy, Smith, & Ogilvie, 1966; Wilson, Wiebe,

& Hoffmann, 2009) that can be applied for sentiment classification. However, these resources may not be appropriated for specific domain and context classification. For example, the word "long" can have many sentiment orientations (e.g. "long battery life", "long debt list", "long Google stocks").

A possible solution is the creation of an opinion lexicon. However, it can be very laborious or even impracticable to construct a comprehensive set of domain opinion words and to determine their orientations.

2.5. Opinion Mining (OM)

OM is the computational treatment of opinion, sentiment and subjectivity in text. This field of study is also frequently defined as "Sentiment Analysis" (Pang & Lee, 2008). OM systems aims to extract "people's opinions, appraisals, attitudes, and emotions toward entities, individuals, issues, events, topics and their attributes" (Liu & Zhang, 2012).

Knowing people's opinion has always been an important piece of information for decision-making processes (Pang & 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 customers may also want to find the opinions of actual customers before they purchase a product.

Social Media platforms (e.g., reviews, forums, blogs and social networks) have enabled an explosion of contents containing opinions regarding several topics. The huge amount of opinionated text in these platforms is a valuable source of opinions of a representative community of users, very useful for organizations and individuals. The extraction of these opinions is extremely difficult for humans. The identification and summarization of important information in large quantities of data is very challenging for the average reader (Liu & Zhang, 2012). These limitations can be overcome by OM systems that mine large amounts of opinionated

contents and automatically extract and summarize the opinions about a topic. These systems have widespread applications, such as:

- Businesses: useful for reputation management, sales prediction, stock management, ads placements, products and services benchmarking.
- Individuals: may support decisions about product purchases,
- Politics: permits to understand what voters are thinking about several political issues such as politicians or political proposals.

In the following subsections, we present an opinion definition that supports OM systems and describe the most common OM tasks.

2.5.1. Opinion Definition

Opinions can be expressed using subjective or objective sentences (Pang & Lee, 2008). Subjective sentences contain some emotional expressions (e.g. "I love that phone"), while objective sentences may present facts denoting opinions (e.g. "The voice of this phone is clear"). Both sentence types can be used to OM.

An opinion may have several characteristics. However, there are five features that provide the necessary information for most of the subsequent analysis. These characteristics are (Liu & 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.
- 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 different strength/intensity levels.
- Opinion Holder: The entity that has expressed the opinion.
- Time: The time when the opinion was expressed.

For example, in the sentence "The new iPad's processor is fast.", the opinion 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 from unstructured text to structured data and permit a more complete and effective knowledge discovery phase.

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.

2.5.2. Document-level Sentiment Classification

This task considers the whole document as the information unit and attributes it a sentiment value (e.g. positive, negative, neutral, rating score). Document-level classification assumes that the document expresses opinions on a single entity from a single opinion holder. It is not appropriated to evaluate and compare diverse entities neither to identify the sentiment regarding multiple entities and aspects mentioned in the document (Liu & Zhang, 2012).

Customer reviews are suited to this type of classification because they are generally written by a single author and are about a single item. However, in many social media contents (e.g. forum, blog), the author may express opinions on multiple entities. reducing the validity of this methodology.

Most existing techniques for document-level sentiment classification are based on supervised learning. These sentiment classifiers are created from classified data using supervised learning methods (e.g. SVM) in order to attribute the correct sentiment class (e.g. positive, negative, neutral). Product reviews provide an useful and common data source for supervised learning because each review usually has a rating assigned by its own author.

2.5.3. Sentence-level Sentiment Classification

Sentence-level sentiment classification is the task of assigning a sentiment class to individual sentences. Document-level sentiment

classification is too coarse for most applications. Thus, the sentence level classification may provide a more adequate analysis 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. Sentence level is appropriate for simple sentences that contain a single opinion because it does not allow to classify more than one element of the sentence (Liu & Zhang, 2012). Many sentiment classification techniques applied in document-level can also be used in this task.

2.5.4. Aspect-level Sentiment Classification

In many situations, the document level and sentence level classification do not supply the necessary detail. A positive document or sentence does not mean that the author has positive opinions on all mentioned entities or aspects. This detailed information could be extremely valuable for decision-making.

Aspect-based sentiment analysis aims 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 & Zhang, 2012). It is a more complex task that requires deeper NLP capabilities to create a richer set of results. Many techniques applied in IE (Freitag & McCallum, 2000; Jakob, 2010; Jin & Ho, 2009; Lafferty, McCallum, & Pereira, 2001), topic modeling (Brody, 2010; Lin & He, 2009; Titov & McDonald, 2008) or clustering (Su et al., 2008) can be used in the extraction of these elements.

2.5.5. Mining Comparative Opinions

Comparing two or more entities is an usual form of sentiment evaluation. These comparative opinions are different from regular opinions, because they involve more entities and have other semantic meanings and syntactic forms. The set of five features that describe regular opinions is insufficient to characterize comparative opinions. These opinions should

be defined by six features: entity 1, entity 2, shared aspects, preferred entity, opinion holder and time (Liu & Zhang, 2012). Thus, the OM process for comparative opinion is different from the process applied to mine regular opinions.

Comparative sentences may have different structures. There are four main comparisons types:

1. Non-equal gradable comparisons (e.g. “X is faster than Y”).
2. Equative comparisons (e.g. “the performance of X is identical to Y.”).
3. Superlative comparisons (e.g. “X is the fastest”).
4. Non-gradable comparisons, comparing but not grading aspects of two or more entities (e.g. "X tastes differently from Y.").

OM should be able to identify the type of the comparative sentence, to extract the comparative relations accordingly and to determine the preferred entity set (Ganapathibhotla & Liu, 2008; Jindal & Liu, 2006). The tasks of extracting entities, aspects, opinion holders and times may be similar to those executed for mining regular opinions. However, mining comparative opinions has a distinct task. It should identify the preferred group of entities among the various entities referred in the sentence.

2.6. Utilization of Microblogging Data to Model and Forecast Stock Market Variables

In this section, we describe the main concepts and research work about mining microblogging data to model and forecast stock market behavior. We included an additional subsection reporting the state of the art about the utilization of similar web data sources. The final subsection summarize the research results and identifies research opportunities.

2.6.1. Investor Sentiment and Attention

Despite, the Efficient Market Hypothesis defend that investors act as rational agents and all existing information is reflected immediately in stock prices, financial research has shown that individual investors systematically deviate from optimal trading behavior (Barber & Odean, 2008; Daniel, Hirshleifer, & Teoh, 2002) and financial decisions are significantly driven by emotion and mood (Nofsinger, 2005). Thus, investor attention and sentiment can be influential in financial decision-making and their measures can be used to identify and exploit stock mispricing.

Investor attention is a limited resource that can have an effect on asset prices and dynamics (Hirshleifer & Teoh, 2003; Merton, 1987). If decision making processes are affected by emotions (Peterson, 2007) one can argue also that investor collective sentiment can also play a role on their investment decisions and, consequently, influence stock market returns and their dynamics.

Microblogging data can be a valuable source to predict these indicators. Investors are increasingly using microblogging services (e.g. Twitter) to express their opinion and share useful information regarding several financial issues. The analysis of these contents can allow the extraction of important signals of sentiment from investors that can contain important information about stock market behavior.

2.6.2. Microblogging data

The most popular microblogging service is Twitter (www.twitter.com). Nevertheless, there is a more specific microblogging platform exclusively dedicated to stock market. StockTwits (www.stocktwits.com) has started in October 2008 and already has more than 200,000 users that share information about the market and individual stocks. Similarly to Twitter, messages are limited to 140 characters and consist of ideas, links, charts and other data.

Microblogging data has distinguishing characteristics that may benefit the creation of sentiment indicators, such as:

- The character constraints require greater objectivity from the author and permits a more accurate linguistic analysis.
- The usage of cashtags (a hashtag identifier for financial stocks) can make it a less noisy source of data.
- Users post very frequently, reacting to events in real-time. This regularity allows a real-time sentiment assessment that can be exploited during the trading day.

However, these contents present some challenges to an adequate sentiment analysis. The short length of messages can generate a lack of contextual information while the informal writing style usually produces many grammatical errors, reducing the performance of NLP resources.

Regarding investor attention, the number of tweets related to a stock can constitute a more precise measurement than traditional sources. Investor attention has been analyzed by the supply side of news, assuming that supply of news is positively correlated with investor attention. Nevertheless, this is not only an imperfect measure of attention but also a limited one, since news availability does not correlate directly with investors' attention to these news. However, when a tweet contains a reference to a stock ticker, it surely indicates that its author is paying attention to that stock.

Microblogging data is usually abundant and readily available at low cost. Thus, the creation of these new indicators may be more rapid, accurate and cost effective than traditional forms (e.g. large-scale surveys). They also might constitute good substitutes for traditional sources, which is particularly relevant where they are not available. In these cases, the proposed indicators can be proxies for such measures.

2.6.3. Stock Market Variables

Stock market variables measure diverse important aspects related to stock market and are influential in investment decisions. The most studied stock market variables in this research topic are:

Returns: Market returns measure changes in the asset value. A common formula to calculate returns is:

$$r_t = \left(P_t - P_{t-1}\right) / P_{t-1} \quad (2)$$

where P_t is the adjusted close price of day *t* and P_{t-1} is the adjusted close price of the preceding day. Adjusted close price is the official closing price adjusted for capital actions and dividends. Returns provide useful information about the probability distribution of asset prices. This is essential for investors and portfolio managers as they use this information to value assets and manage their risk exposure.

Volatility: Volatility is a latent measure of total risk associated with a given investment. Volatility can be estimated using different approaches. Estimates of volatility are essential for portfolio selection, financial assets valuation and risk management.

Trading volume: Trading volume is the number of shares traded in each day during a trading session. Volume can be used to measure stock liquidity, which in turn has been shown to be useful in asset pricing as several theoretical and empirical studies have identified a liquidity premium. Liquidity can help to explain the cross-section of expected returns (Amihud, Mendelson, & Pedersen, 2005).

2.6.4. Related work using microblogging data for stock market prediction

Microblogging data has motivated some studies about its relationship with stock market variables. Sentiment indicators extracted from microblogging data had predictive value for future stock price directions (Bollen et al., 2011; Oh & Sheng, 2011) and returns (Mao et al., 2011; Sprenger &

Welpe, 2010) and were correlated with volatility (Fuehres et al., 2011). Trading volume were also correlated with the number of tweets (Sprenger & Welpe, 2010).

Bollen et al. (Bollen et al., 2011) measured collective mood states (e.g. "positive", "negative", "calm") through sentiment analysis applied to large scale Twitter data, although tweets were related with generic sentiment (e.g. "I'm feeling") and not directly related to stock market. These messages were classified using two different lexicons: the MPQA Subjectivity Lexicon (MPQA) (Wilson et al., 2009) and GPOMS, a created lexicon based on an existing psychometric instrument. Applying a Self-organizing Fuzzy NN, they found an accuracy of 86.7% in the prediction of the Dow Jones Industrial Average daily directions and a substantial reduction in the Mean Average Percentage Error (MAPE).

Sprenger and Welpe (Sprenger & Welpe, 2010) have used sentiment analysis on stock related tweets collected during a 6-month period. To reduce noise, they selected Twitter messages containing cashtags of S&P 100 companies. Each message was classified by a Naïve Bayesian method trained with a set of 2,500 tweets. Results showed that sentiment indicators are associated with abnormal returns and message volume is correlated with trading volume.

Mao et al. (Mao et al., 2011) surveyed a variety of web data sources (Twitter, news headlines and Google search queries) and tested two sentiment analysis methods to predict stock market behavior. They used a random sample of all public tweets and defined a tweet as bullish or bearish only if it contained the terms "bullish" or "bearish". They showed that their Twitter sentiment indicator and the frequency of occurrence of financial terms on Twitter are statistically significant predictors of daily market returns.

Oh and Sheng (Oh & Sheng, 2011) resorted to a microblogging service exclusively dedicated to stock market. They collected 72,221 micro blog postings from StockTwits about 1,909 stocks from NASDAQ and NYSE,

over a period of three months. The sentiment of the messages was classified by a BOW approach that applies a J48 classifier to produce a learning model. They verified that the extracted sentiment appears to have strong predictive value for future stock price directions.

Ruiz et al. (Ruiz et al., 2012) verified whether microblogging features are correlated with stock prices and trading volume. They collected Twitter data about for 150 companies of the S&P 500 index for the first half of 2010. The used microblogging features are related to its activity (e.g., number of posts, number of re-posts) or measure properties of an induced interaction graph (e.g., number of connected components, statistics on the degree distribution). Results show that trading volume is correlated with some graph-based features but the stock price is not strongly correlated with any of the microblogging features. Nevertheless, they created a simulator of investments applying a Twitter-Augmented Regression model that outperformed other baseline strategies.

Fuehres et al. (2011) (Fuehres et al., 2011) analyzed six months of random twitter feeds and created daily mood indicators by counting all tweets containing some key words (e.g., "fear", "worry", "hope"). They found that these emotional indicators are significantly negatively correlated with Dow Jones, NASDAQ and S&P 500, but are positively correlated with VIX.

Figure 2 illustrate the research results and Table 3 summarize these studies.

2.6.5. Related work applying other sources of web social data

Various sources of web social data have been used to model and predict stock market variables. Social media services, such as blogs and message boards, have similar characteristics to microblogs. For instance, they are an abundant source of opinionated text contents created in an interactive environment and using an informal writing style. The number of microblogging messages is also comparable to internet searches.

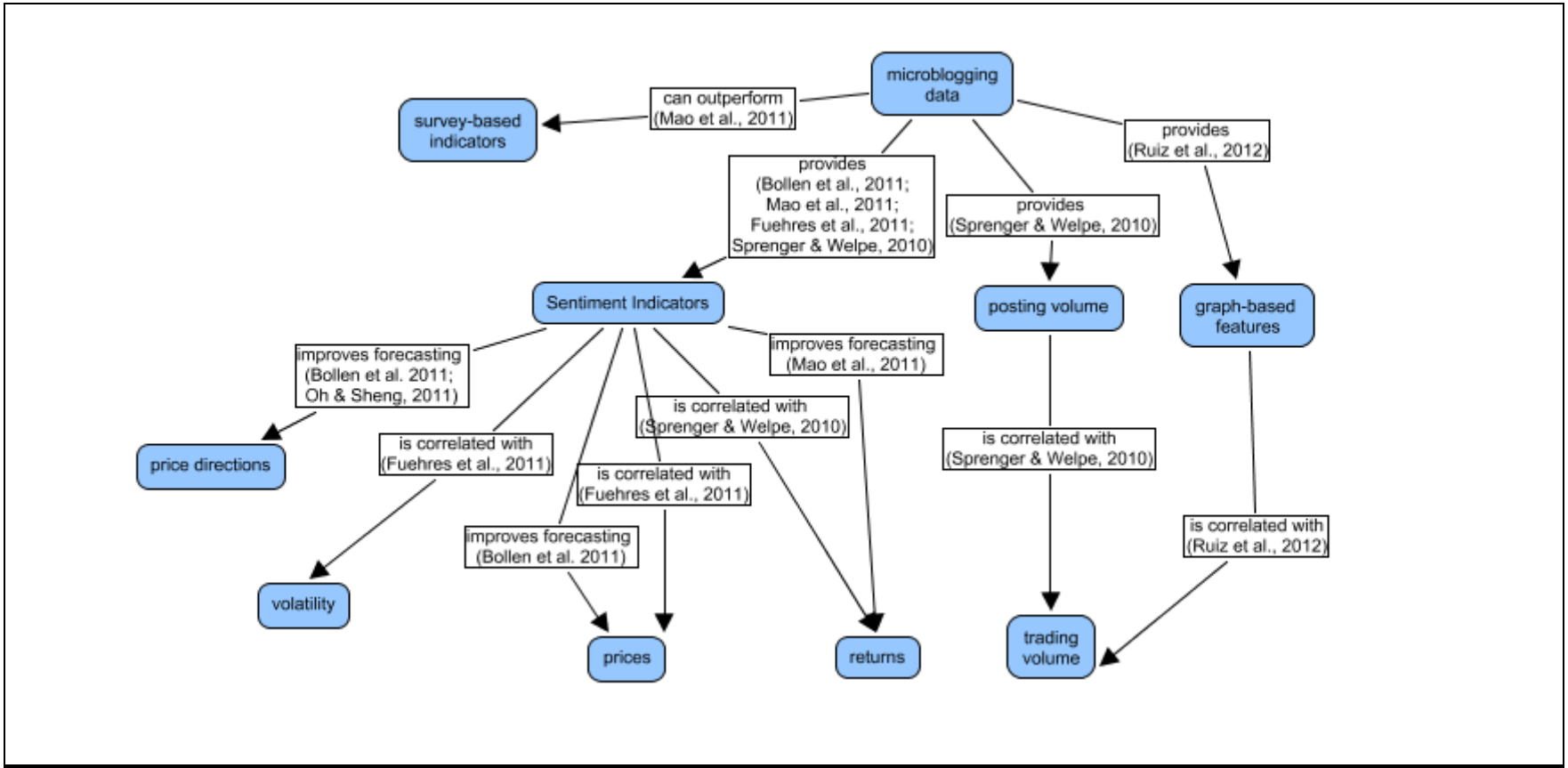


Figure 2. Literature Map About Mining Microblogging Data to Model and Forecast Stock Market Behavior

Table 3. Research about Mining Microblogging Data to Model and Forecast Stock Market Behavior

Paper	Microblogging data collected	Sentiment Analysis Method	Analysis of the relationship of microblogging data and stock market variables					
			Microblogging features	Stock Market variables	Methods	Test Period	Evaluation	Results
Sprenger and Welpe (2010)	Twitter data containing cashtags of S&P 100 companies, from January 1 to June 30, 2010.	Naïve Bayesian method trained with 2,500 tweets to classify messages as "buy","hold" or "sell"	Bullishness index, message volume, disagreement index	Abnormal returns, trading volume and volatility	Pairwise correlations, contemporaneous regressions and time-sequencing regressions	–	Correlation coefficients, regression coefficients, R2 and F-value	Bullishness index is correlated with abnormal returns; message volume is correlated with trading volume
Bollen et al. (2011)	Twitter data containing specific sentiment expressions, from February 28 to December 19, 2008.	Messages classified using two lexicons: MPQA and GPOMS	6 mood indicators (Calm, Alert, Sure, Vital, Kind, Happy) and a generic sentiment indicator	DJIA prices and price movements	Granger Causality Analysis, Self-organizing Fuzzy NN	From December 1 to December 19, 2008	MAPE and direction accuracy	86.7% accuracy in predicting DJIA directions; less 15% MAPE than baseline method
Mao et al. (2011)	Random sample of Twitter data, from July 2010 to September 2011.	Tweets classified as bullish or bearish if contains the terms "bullish" or "bearish"; Tweet volumes of financial terms	Twitter Investor Sentiment; Tweet Volumes of Financial Search Terms	DJIA price, returns, trading volumes, market volatility (VIX)	Correlation Analysis, Granger Causality Analysis, Multiple Regression Analysis	From August 31 to September 29, 2011	R^2, MAPE and direction accuracy	Twitter sentiment and Tweet volumes of financial terms are statistically significant predictors of daily market returns.
Oh and Sheng (2011)	StockTwits data containing tickers of 1,909 stocks (NASDAQ and NYSE), from May 11 to August 8, 2010.	J48 classifier trained with 7109 messages that labels messages as "Bullish", "Bearish" ou "Neutral"	Bullishness Index; posting volume; author information	Stock price directions	Multiple Regression Analysis	From August 1 to August 8, 2010	Direction accuracy	Bullishness index appears to have strong predictive value for future stock price directions
Fuehres et al. (2011)	Random sample of Twitter data, from March 30 to September 7, 2009.	Daily mood indicators created by counting all tweets containing some key words (e.g., fear, worry, hope)	Mood indicators	Index values of Dow Jones, NASDAQ, S&P 500, VIX	Correlations	–	Correlation coefficients	Mood indicators are significantly negatively correlated with Dow Jones, NASDAQ and S&P 500, but are positively correlated with VIX
Ruiz et al. (2012)	Twitter data containing expressions related to 150 companies in the S&P 500 index, during the first half of 2010.	–	Activity features; Graph-based features	Stock prices and trading volume	Correlations	–	Correlation coefficients	Trading volume is correlated with some graph-based features

Therefore, research using these resembling data sources can provide important information to this topic.

In these subsection, we will describe research work applying internet searches, blog systems and message boards. Figure 3 illustrate research results using diverse web data sources.

2.6.5.1. Internet Searches

The search frequency in search engines is considered a direct measure of investor attention. When someone searches for a stock is assuredly paying attention to it and internet users usually use a search engine to collect information. The search volume reported by Google is likely to be representative of the internet search behavior of the general population because represents approximately 70% of worldwide internet searches[2]. Therefore, aggregate search frequency in Google is a direct and unambiguous measure of attention that can be processed in a timely fashion (Da, Engelberg, & Gao, 2011).

Since 2004, Google provides data about searches carried out in this search engine. Google Trends (http://www.google.com/trends/) analyzes a portion of Google web searches to compute how many searches have been done for the terms entered, relative to the total number of searches done on Google over time. The information provided by Trends is updated daily and results are normalized. Google Trends also allows more advanced queries, permitting to compare search volume patterns across specific regions, categories, time frames and properties.

Google queries are considered a more direct and accurate measure than other attention measures, such as the number of news available or advertising expenditures (Da et al., 2011; Ding & Hou, 2011). These traditional measures do not assure that investors are paying attention on them.

[2] http://www.netmarketshare.com/search-engine-market-share.aspx?qprid=4&qpcustomd=0

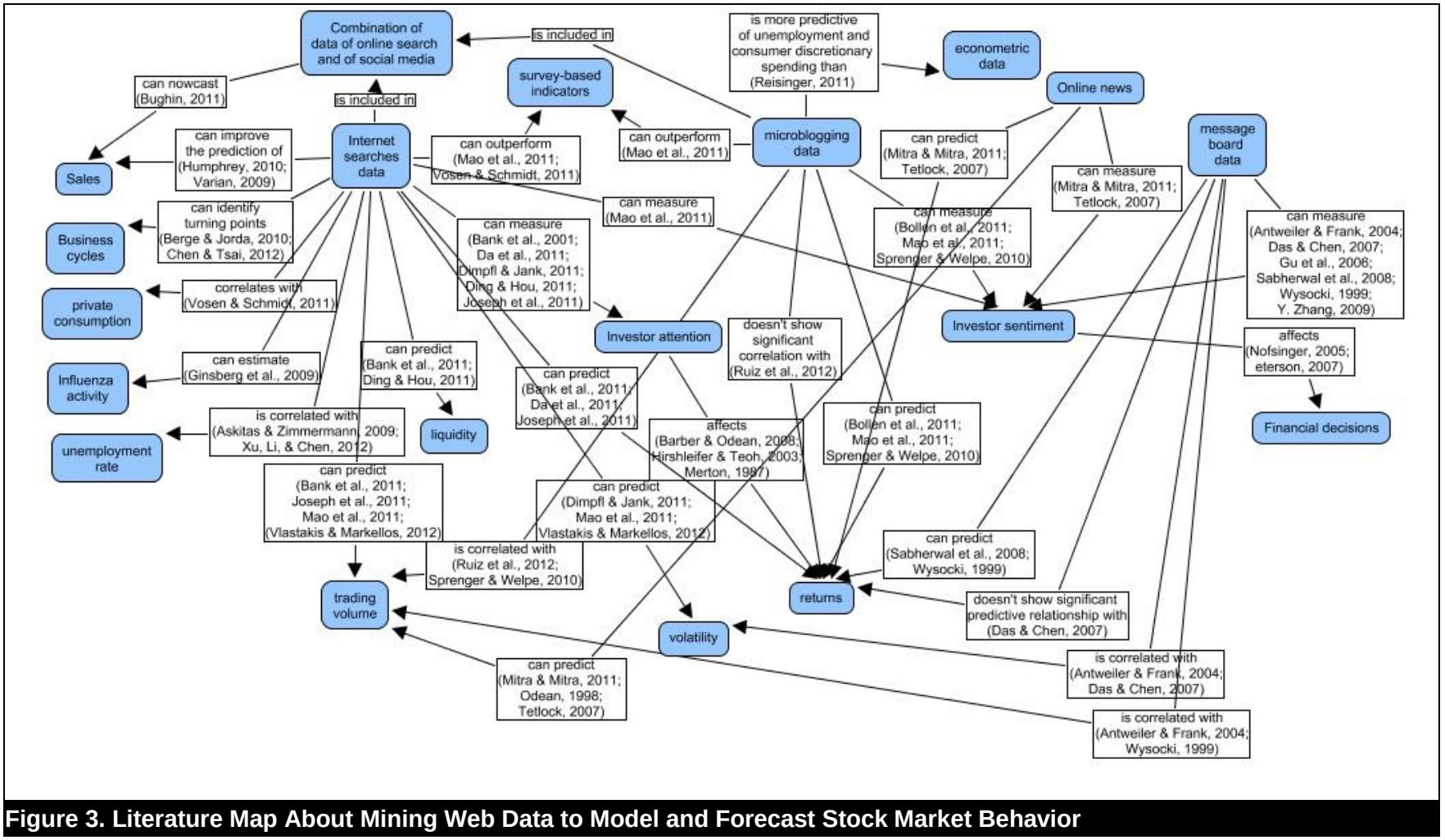


Figure 3. Literature Map About Mining Web Data to Model and Forecast Stock Market Behavior

The stock ticker search is a common search strategy used (Da et al., 2011; Ding & Hou, 2011; Joseph, Wintoki, & Zhang, 2011) because the effort required to process results of a ticker query is considered to be only worthwhile for someone who is seriously considering an investment decision. It is unlikely that someone searches for a company's ticker for other reasons (Joseph et al., 2011). The search query for a ticker symbol is also likely to characterize the behavior of retail investors (Bank, Larch, & Peter, 2011; Da et al., 2011; Joseph et al., 2011). Institutional investors can easily access more sophisticated and expensive information databases.

Besides improving predictions accuracy, search data can provide faster information. Unlike common economic variables, query logs are readily available through online Web services, so information can be delivered in a timely manner. Several studies have demonstrated the applicability of search data to predict or to real-time assess some stock market variables.

An increase in search queries is consistently associated with a rise in trading activity. Thus, search intensity can, reliably and timely, predict trading volumes (Bank et al., 2011; Joseph et al., 2011; Mao et al., 2011), and this relationship becomes stronger during "high return" market states (Vlastakis & Markellos, 2012).

The volume of searches seems to be correlated with the stock market volatility and have predictive power for future volatility. Dimpfl and Jank (2011) studied the dynamics of stock market volatility and retail investor attention measured by internet search queries. They found that search queries are particularly useful to predict volatility in high-volatility phases. In a long-run variance decomposition, log search queries accounted for 9% to 23% of the variance of log stock market volatility (Dimpfl & Jank, 2011). Mao et al. (2011) also found that there is a significant correlation between weekly Google search volumes of financial terms and market volatility. Vlastakis and Markellos (2012) have shown that variations in information demand appear to have a significant effect at the individual stock and overall market level in terms of historical volatility.

Higher Internet search volume is also associated to higher levels of stock liquidity. More searches of stock ticker from Google Trends represent increased retail attention. Consequently, that stock is associated with a larger shareholder base, and with significantly improved stock liquidity (Ding & Hou, 2011). This is also consistent with the "investor recognition hypothesis". Although investors continuously receive information, they are unlikely to pay the same level of attention on each piece of the information, because there are limits on the central cognitive-processing capacity of human beings. In markets with incomplete information, information asymmetry becomes more severe for stocks with lower investor recognition. When individual investors pay more attention to a stock by actively searching it on the internet, they acquire relevant information mitigating information asymmetry and therefore they are more likely to invest in them. As a result, stock liquidity increases with the active attention of retail investors (Bank et al., 2011; Ding & Hou, 2011).

Search data can also improve the prediction of stock market returns. An increase in search volume is associated with temporarily higher future returns (Bank et al., 2011; Da et al., 2011). This confirms the attention-induced price pressure hypothesis of Barber and Odean (2008) because, in these studies, a positive short-run relationship between changes in Google search volume and future stock returns is reversed on the long-run. They attribute it to buying pressure from uninformed investors, which is consistent to Google search volume being a proxy for retail attention. Joseph et al. (2011) besides confirming that online search intensity reliably predicts abnormal stock returns, also found that the sensitivity of returns to search intensity is positively related to the difficulty of a stock being arbitraged.

2.6.5.2. Blogs

Blogs are websites that provide information created, mostly, by individuals. This content is displayed in reverse chronological order and can be about any topic including finance. They are usually interactive, enabling visitors to comment each post. Blog contents typically have less

quality than other sources such as newspapers or magazines. It is less edited and it is more suitable to rumors. However, information distributed in blogs carries many different perspectives and insights of many people (Leary, 2011). This potentially offers access to important information.

There are not many studies applying blogging data to model and forecast stock market behavior. Choudhury et al. (De Choudhury, Sundaram, John, & Seligmann, 2008) studied the correlation between contextual features of the blogosphere and stock market movements in four technological companies. Several contextual features were extracted such as the number of posts, the number of comments, the length and response time of comments, strength of comments and the different information roles played by users (e.g., early responders, loyals, outliers). This study achieved approximately 78% accuracy in predicting the weekly magnitude of movement and 87% for the weekly direction of movement.

Kharratzadeh et al. (Kharratzadeh & Coates, 2011) identified groups of companies whose stock prices are more likely to be correlated in the future. They applied clustering methods using blogging data and historical stock prices. The similarity measure between two companies was the number of mutual appearances of the company names in blog articles in a given time period.

Yu et al. (Yu, Duan, & Cao, 2013) analyzed the impact of diverse social media (e.g., blogs, Twitter, forums) and conventional media on stock market performance (Google News). Regarding blogging data, they analyzed the sentiment of contents from Google Blogs and extracted four daily indicators for each company: the number of positive sentiment blogs; the number of negative sentiment blogs; the total number of mentions in blogs; overall sentiment in blogs. They found that blog sentiment indicators have a positive impact on return and on risk.

2.6.5.3. Message Boards

A message board is a discussion website where users can post messages about a specific subject that will be available to be read and

replied by other users. It has a hierarchical structure, being composed by subforums, topics, threads and replies.

There are several messages boards dedicated to stock market (e.g., finance.yahoo.com, ragingbull.com, thelion.com). The growth of the World Wide Web and on-line stock trading has greatly increased the popularity of virtual investor forums (Wysocki, 1999).

The message content is diverse. Many postings are pure noise, with only limited connection to the stock. However, other messages can contain important information. There are sophisticated debates about company financial disclosures with investors dissecting, interpreting, and debating financial reports. Some message postings can even contain relevant information not available from other public sources. Additionally, it can be a barometer of investor opinion about several stock market issues (Wysocki, 1999). Thus, this source can provide valuable information in a timely manner.

The analysis has been made by three modes: message volume (Antweiler & Frank, 2004; Das & Chen, 2007; Sabherwal, Sarkar, & Zhang, 2008; Wysocki, 1999), message sentiment (Antweiler & Frank, 2004; Das & Chen, 2007) or reputation of the message poster (Gu, Konana, Liu, Rajagopalan, & Ghosh, 2006; Zhang, 2009). The reputation of the poster seems to be helpful to determine sentiment and can complement message content analysis. It can be used to decide whether and how to utilize poster's comments in formulas or algorithms when determining sentiment (Gu et al., 2006; Zhang, 2009)

Some studies found correlations between message board data and stock return (Antweiler & Frank, 2004; Sabherwal et al., 2008; Wysocki, 1999). Sabherwal et al. (2008) revealed that posting volume positively correlates with stock's abnormal return on the same day and also predict next day's abnormal returns. Wysoki (1999) showed that a tenfold increase in message postings in the overnight hours led to a 0.7% increase in next day stock returns. Antweiler & Frank (2004) verified that higher negative

postings help predict negative subsequent returns. However, Das and Chen (2007) did not find significant predictive relationship between sentiment and stock prices.

The trading volume is also associated to message board volume (Antweiler & Frank, 2004; Wysocki, 1999). Wysoki (1999) found a strong positive correlation between the volume of messages posted on the discussion boards during the hours that the stock market is closed and the next trading day's volume. A tenfold increase in message volume signified an average increase in the next day's stock grade volume of approximately 15.6%.

A significant correlation between posting volume and volatility was also identified in some studies (Antweiler & Frank, 2004; Das & Chen, 2007).

2.6.6. Summary

Research studies has presented very interesting results about the utilization of microblogging data to model and forecast stock market variables. In these studies, sentiment and attention indicators created using microblogging data had valuable information to predict stock price directions (Bollen et al., 2011; Oh & Sheng, 2011) and returns (Mao et al., 2011; Sprenger & Welpe, 2010) and were correlated with volatility (Fuehres et al., 2011) and trading volume (Sprenger & Welpe, 2010).

Notwithstanding, these results need to be interpreted with caution because most of these studies do not perform a robust evaluation. For instance, only correlation analysis was performed in (Fuehres et al., 2011; Ruiz et al., 2012), only modeling (and not prediction) was addressed in (Sprenger & Welpe, 2010), and very short test periods were applied in (Bollen et al., 2011) (19 predictions), (Mao et al., 2011) (20 and 30 predictions) and (Oh & Sheng, 2011) (8 predictions). Moreover, the majority of these studies analyzed the relationship with prices or prices movements, rather than returns. However, forecasting stock returns is more difficult and add more information to investment decisions.

This is a very recent topic with many research opportunities to improve results. For instance, the state of the art TM models have not been applied in this domain, namely in the creation of sentiment indicators. The applied sentiment analysis algorithms utilize BOW methods that do not allow a rigorous textual analysis. Sentiment classification has been performed at the sentence level, labeling the sentiment of the whole sentence and ignoring the opinion about specific important stock features (e.g., financial parameters, product performance). Moreover, some microblogging characteristics (e.g., profusion of grammatical errors, specific structure and terminology) influence the performance of NLP resources and have not been properly addressed. Additionally, stock market has a distinct vocabulary that has different meaning than common language (e.g., "long", "short", "bull", "bear"). A proper contextual analysis should consider this specific terminology. The utilization of more advanced algorithms, adapted to social media and stock market conversations, could produce more accurate and complete sentiment indicators (e.g. sentiment scores for diverse stock features) that can potentially improve the forecasting of stock market behavior.

Additionally, the minimum periodicity used in this topic has been the daily frequency. However, the utilization of a timely analysis, performed quickly after posting (e.g. minutes or hours) may enable a better exploitation of sentiment effects in stock market.

Few results confirm the ability of microblogging features to predict trading volume and volatility. However, research carried out with similar data sources (e.g., message boards, blogs, Google searches) has frequently created indicators correlated with trading volume (Antweiler & Frank, 2004; Bank et al., 2011; Joseph et al., 2011; Mao et al., 2011; Vlastakis & Markellos, 2012; Wysocki, 1999) and volatility (Antweiler & Frank, 2004; Das & Chen, 2007; Dimpfl & Jank, 2011; Mao et al., 2011; Vlastakis & Markellos, 2012). Therefore, it may be productive to further investigate the capacity of microblogging data to predict these stock market variables.

3. Experiments on Modeling Stock Market Behavior Using Investor Sentiment Analysis and Posting Volume from Twitter

3.1. Introduction

The assessment of the state of the art revealed that mining microblogging data may provide measures of investor sentiment and attention that are more accurate, and supplied in a more timely and cost effective manner than traditional ones. Research studies had shown that microblogging features had important information to model and predict stock price directions (Bollen et al., 2011; Oh & Sheng, 2011), returns (Mao et al., 2011; Sprenger & Welpe, 2010), volatility (Fuehres et al., 2011) and trading volume (Sprenger & Welpe, 2010).

However, this research topic is still in its infancy and research results are insufficient and inconclusive. For instance, few studies confirmed the relationship between microblogging features and trading volume or volatility. The analysis of returns was neglected in favor of less informative variables such as stock prices or price movements.

In this project, we intend to extend research in this topic by focusing in the mentioned stock market variables, by testing different sentiment analysis methods and by concentrating in a specific sector. We utilized 32 days of Twitter and stock market data to model the next day returns, volatility and trading volume of nine major technological companies. We focused on a highly posted sector (i.e. technological sector) that may provide more representative indicators of investor sentiment and attention. The sentiment indicators were created by exploring five popular lexical resources and two new proposed lexicons: emoticons; and ALL, which joins the six remaining resources. The application and combination of broad and highly experimented lexical resources can produce more

precise measures of investor sentiment. Moreover, many microblogging authors use emoticons to indicate their dominant sentiment.

A detailed description of the materials, methodology and results of this project are presented in the remaining sections.

3.2. Materials and Methods

In this project, we evaluated the informative content of posting volume and sentiment indicators extracted from Twitter data to model next day returns, volatility and trading volume of nine technological companies: AMD (AMD), Amazon (AMZN), Dell (DELL), Ebay (EBAY), HP (HPQ), Google (GOOG), IBM (IBM), Intel (INTC) and Microsoft (MSFT).
This section describes the data sources, sentiment analysis methods, regression models and evaluation procedures executed in order to achieve the proposed research objectives. The figure 4 illustrates the whole process. All tasks were performed using the open source R tool (Team, 2012) running on Linux server.

3.2.1. Twitter Data

Twitter is by far the most popular microblogging service. Recently, this platform defined a specific term (cashtag) to identify conversations about a specific stock. Cashtags are composed by the company ticker preceded by the "$" symbol (e.g., $IBM). These symbols are commonly used by the investor community in discussions related to the respective company. Concentrating on only these messages reduces the amount of irrelevant messages, resulting in a less noisy data set.
We collected data about each technological company: AMD, Amazon, Dell, Ebay, HP, Google, IBM, Intel and Microsoft. These companies were chosen because they belong to a sector that has a substantial posting volume on Twitter and therefore can be indicative of investors' level of attention on these stocks.

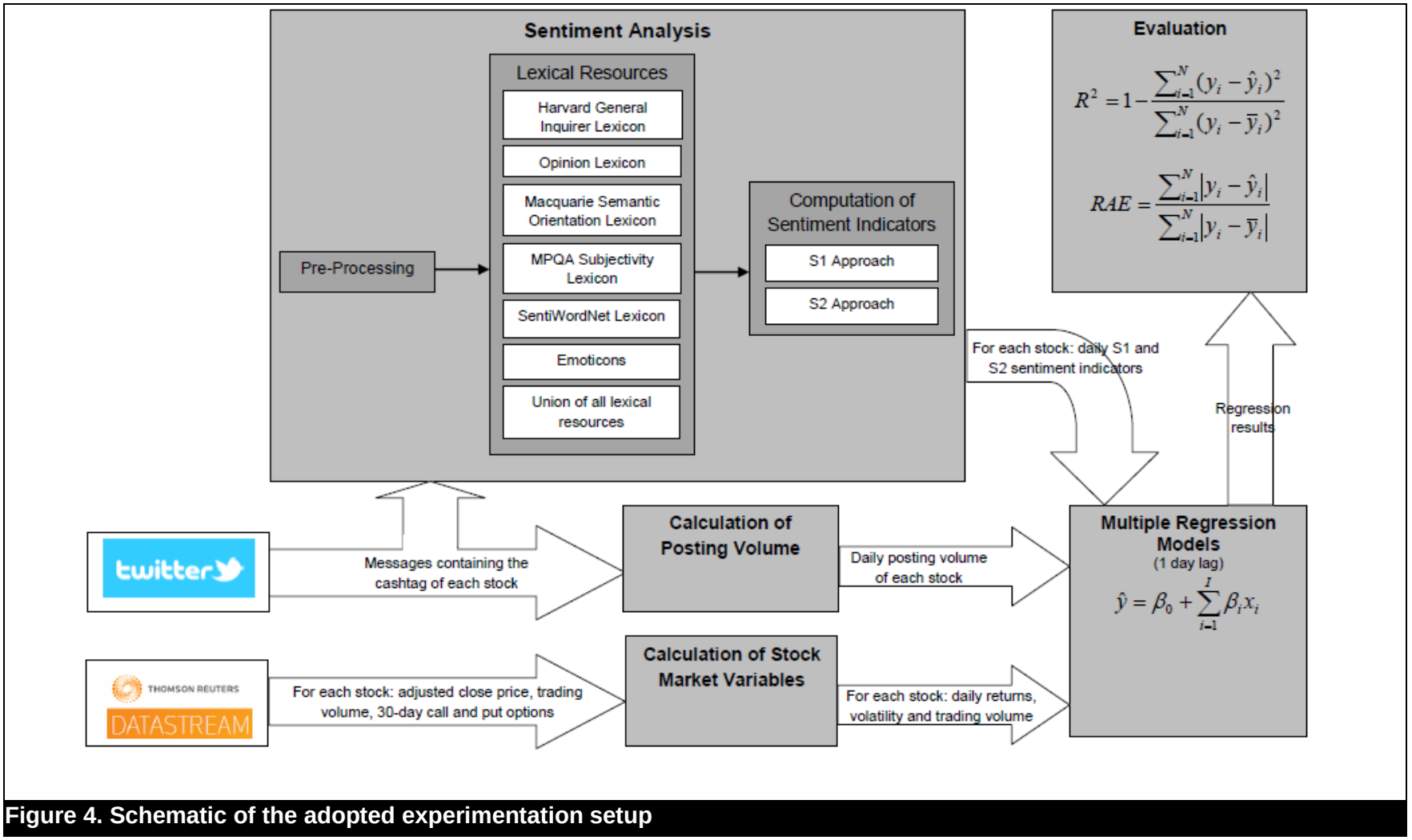


Figure 4. Schematic of the adopted experimentation setup

For each company, we collected Twitter data, on a daily basis (considering working days) from December 24, 2012 to February 8, 2013. Figure 5 plots the total number of tweets collected for each technological company.

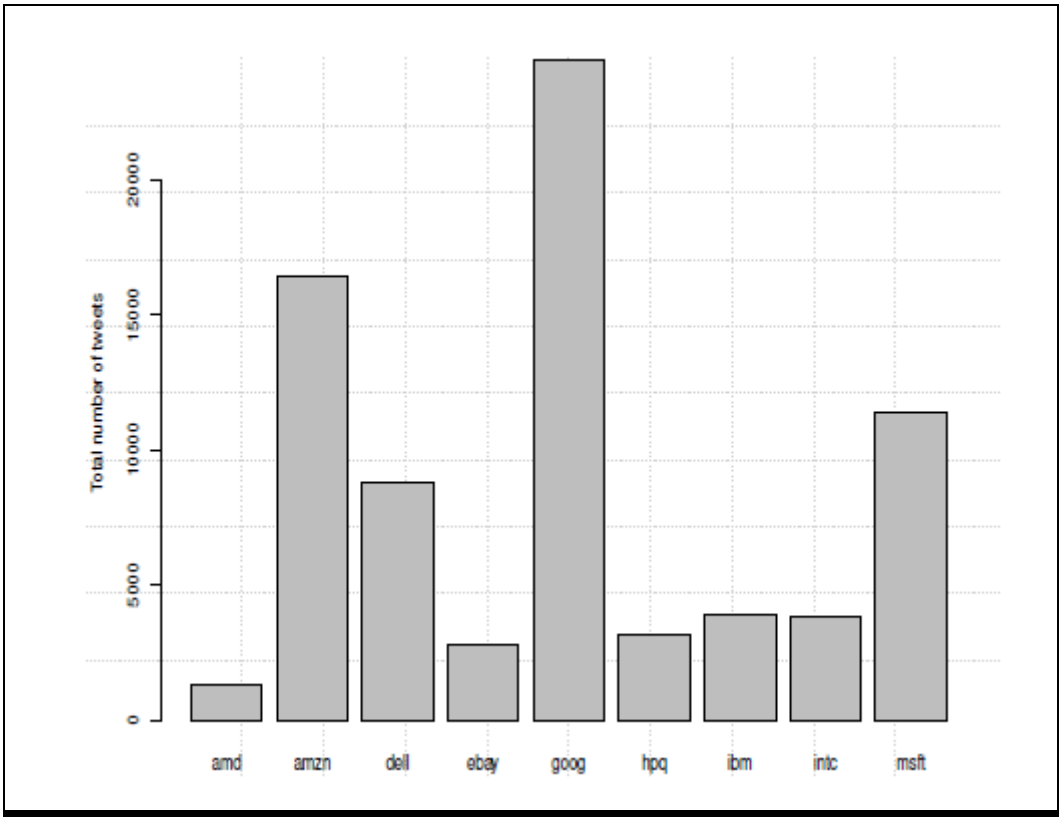


Figure 5. Total number of tweets collected for the nine selected technological companies

All daily tweets have been gathered using the Twitter REST API[3] and applying queries containing cashtags of all studied stocks (i.e. $AMD, $AMZN, $DELL, $EBAY, $GOOG, $HPQ, $IBM, $INTC, $MSFT). All attributes considered to be important for this project (e.g., text, creation date, id) were stored in a MongoDB database. We executed all data requests and operations related to the processing, selection and storage of Twitter data using the R tool and applying many open source packages such as rjson, rmongodb, ROAuth and tm.

[3] https://dev.twitter.com/docs/api

3.2.2. Stock Market Data

The stock market variables analyzed in this study were the daily returns, volatility and trading volume. Returns were calculated using the adjusted close prices collected from Thompson Reuters Datastream and the trading volume indicators were retrieved directly from the same source.
Volatility can be estimated using different approaches. Previous studies have found that the model-free implied volatility index in an appropriate estimator of volatility (G. J. Jiang, 2005). In this project, the average of the implied volatility for a 30-day to maturity Call and Put options contracts for each stock was used to measure volatility.

3.2.3. Sentiment Analysis Methods

In this subsection, we describe the operations and resources needed to create the sentiment indicators, namely the pre-processing tasks, lexical resources and sentiment analysis methods.

3.2.3.1. Pre-processing

Pre-processing is an important phase that allows a better performance of the subsequent TM and DM tasks. We performed usual and simple pre-processing tasks, such as:

- Substitute special characters in HTML (e.g., "&amp", "&quot") using regular expressions to find and replace these characters. These operations are easy and fast to execute and permit to convert HTML expressions to regular text.
- Remove punctuation applying the removePunctuation function included in the tm R package. In these studies, punctuation had less value because we did not perform syntactical analysis.
- Convert all words in lowercase to reduce the number of distinct words. We executed this operation through the tolower R function.
- Perform the tokenization of the messages, in order to divide them into words. These lists of words are the input of the next TM

operations. Tokenization was performed utilizing the function scan_tokenizer included in the tm R package.

- Execute the stemming of the text to decrease the number of words analyzed in the following tasks. This operation reduces each word to its stem (e.g. stem of words "waiting" and "waited" is "wait"). We applied the function wordStem included in the Rstem R package.

3.2.3.2. Lexical Resources

The sentiment analysis methods applied in this project exploited five different and popular lexical resources in order to evaluate the usefulness of each resource, as well as their complementarity. The lexical resources were:

1. Harvard General Inquirer (GI) (Stone et al., 1966): This resource comprise 11788 words classified in 182 categories. These categories come from four sources: the Harvard IV-4 dictionary; the Lasswell value dictionary; categories recently constructed; and "marker" categories containing syntactic and semantic markers. We exploited this resource by producing a list with the "positive" category words and another list with the "negative" category words. The syntactic information was discarded because we did not analyze the text syntactically.
2. Opinion Lexicon (OL) (M. Hu & Liu, 2004): This lexicon contains two lists of positive and negative opinion words for English, including misspelled words that appear frequently in social media contents. We applied this lexicon without any transformation.
3. Macquarie Semantic Orientation Lexicon (MSOL) (Mohammad, Dunne, & Dorr, 2009): This lexicon comprises more than 75 thousand n-grams, labeled as positive or negative.
4. MPQA Subjectivity Lexicon (MPQA) (Wilson et al., 2009): This lexicon is part of OpinionFinder, a system that identifies various aspects of subjectivity (e.g. sources of opinion, sentiment expressions) in text. MPQA Subjectivity Lexicon has more than 8000 entries with the following attributes:

- Type: The word is classified as strongsubj, if it is subjective in most contexts, or weaksubj, if it only has certain subjective usages.
- Len: Refers to the number of words in the entry.
- Word1: Indicates whether contains the word or its stem.
- Pos1: Identifies the POS of the word (i.e. noun, verb, adverb or adjective). It may be anypos, meaning that POS is irrelevant for the polarity.
- Stemmed1: Indicates whether the word is stemmed (y) or not (n).
- Priorpolarity: - Classifies the out of context polarity of the word. It may be positive, negative, both or neutral.

5. SentiWordNet (SWN) 3.0 (Baccianella et al., 2010): It is a lexical resource that assigns, to each synset of WordNet, a positivity and a negativity score, varying from 0 to 1. A synset is a group of words or expressions that are semantically equivalent in some context. Each word may appear multiple times with different scores because it can belong to various synsets of Wordnet. In our study, we used the average positivity and negativity score for each word because we did not analyze the contextual polarity.

We proposed two additional resources, termed Emoticons and ALL. The former is based on the simpler analysis of positive (e.g., ":-)" or ":)") and negative (e.g., ":-(" or ":(") emoticons. If a positive emoticon is present in the text, then we add 1 to the positivity score and similarly we increase (+1) the negative score if a negative emoticon is detected. The latter lexicon merges all 6 previous lexicons (GI, OL, MSOL, MPQA, SWN, Emoticons) by producing a union of all positive, negative and neutral score rules.

3.2.3.3. Sentiment Analysis Approaches

We proposed and explored two simple and global sentiment analysis approaches. In the first approach (S1), we counted the daily total number of words that are considered positive and negative by each lexical resource (total of two sentiment variables). As an example, if a tweet has

2 positive words and 3 negative words, we add 2 to the daily positivity score and 3 to the daily negativity score. In the SWN situation, we added the positivity and negativity score of each word.

Regarding the second sentiment approach (S2), we classified each individual tweet, by considering the "positive" and "negative" words that it contains. A message is considered: positive, if the number of "positive" words is higher than the number of "negative" words; negative, if the number of "negative" words is higher than the number of "positive" words; and neutral, if the number of both word polarity types is equal. In the SWN approach, we compared the total positivity and total negativity score for each tweet. The total of sensitive variables measured using each lexical resource is thus two for S1 (total number of positive and negative words) and three for S2 (total number of positive, neutral and negative classified tweets).

The rational for this choice is that the proposed approaches are very easy to implement and test. For example, a tweet that contains "I do not really like $GOOG prices" will have a neutral effect under both proposed strategies and it is not wrongly classified as positive, while correctly identifying equivalent or even more complex negative posts would require a quite sophisticated parsing. Moreover, since we analyzed a very large number of daily tweets (e.g. several thousands of posts for GOOG, see Figure 5), these simple approaches should produce good global results.

3.2.4. Regression Models

We adopted the multiple regression model for the creation of our predictive models. According to the Occam's razor principle and giving that it has few internal parameters, this model is less prone to overfit the data. Such model is defined by the equation (Hastie, Tibshirani, & Friedman, 2009):

$$\hat{y} = f(x_1,...,x_I) = \beta_0 + \sum_{i=1}^{I} \beta_i x_i \quad (3)$$

where $\hat{y}$ is the predicted value for the dependent variable *y* (target output), x_i are the independent variables (total of *I* inputs) and $\beta_0,...,\beta_i$ are the set of parameters to be adjusted, usually by applying a least squares algorithm. Due to its additive nature, this model is easy to interpret and has been widely used in Finance.

As a baseline method, we adopted a regression model that has one input: the target stock market variable from the previous day (*t*-1). For all metrics, we measure the value of microblogging data if the respective regression model is better than the baseline method. If it happens, we consider that microblogging features add useful information to the model.

Next, we present the regression models applied for each studied stock market variable: returns, trading volume and volatility.

3.2.4.1. Returns

The relationship between the information content of microblogging data and daily returns was tested by regressing the return for each company on several combinations of microblogging variables.

For each stock, we modeled S1 and S2 sentiment variables per lexical resource (GI, OL, MSOL, MPQA, SWN, Emoticons and ALL). For all these models, there is only input x_1 = Pos - Neg, where Pos and Neg denote the positive and negative counts (according to method S1 or S2). The baseline uses the input $x_1 = r_{t-1}$. We also test a regression model that combines the best sentiment variable with the baseline ($x_1 \oplus r_{t-1}$).

3.2.4.2. Trading Volume

Trading volume is usually correlated with investor attention. The number of tweets is the microblogging variable that is more closely related with investors attention. Thus, for each stock, we tested the relationship with trading volume by measuring the regression between trading volume (v_t) and the previous day total number of related tweets (n_{t-1}). The baseline uses only the previous day trading volume (v_{t-1}).

3.2.4.3. Volatility

Research has shown that some measures of investor attention (e.g., Google searches) are correlated with volatility (Antweiler & Frank, 2004; Da et al., 2011; Dimpfl & Jank, 2011). Therefore, we focused on posting volume to assess the informative content of microblogging data to model volatility. For each stock we used the following combination of independent variables:

- previous day number of tweets (n_{t-1});
- previous day volatility (σ_{t-1});
- previous day volatility and number of tweets ($\sigma_{t-1} \oplus n_{t-1}$);
- previous day volatility and number of tweets for the previous two days t-1 and t-2 ($\sigma_{t-1} \oplus n_{t-1} \oplus n_{t-2}$).

3.2.5. Evaluation

To measure the quality of fit of the regression models we used two metrics: coefficient of determination R^2 and Relative Absolute Error (RAE). These are given by (Witten & Frank, 2005):

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2}$$
$$RAE = \frac{\sum_{i=1}^{N}|y_i - \hat{y}_i|}{\sum_{i=1}^{N}|y_i - \bar{y}_i|} \quad (4)$$

where y_i and $\hat{y}_i$ are the target and fitted value for the *i*-th day, *N* is the number of days considered and $\bar{y}_i$ is the average of the target values.

Both metrics are scale independent. The ideal regression will produce a R^2 of 1.0, while an R^2 closer to 0 indicates a bad fit. The lower the RAE, the better the model, where 1.0 means that the regression method has similar performance as the constant average predictor. When compared with RAE, R^2 is more sensitive to high individual errors.

The value of *N* depends on the regression model input variables. If only previous day values are used (*d*-1), *N* corresponds to 31. If a lag of two

days (*d*-2) is included, *N* is equal to 30. For all metrics, we measure the value of Twitter based data if the respective regression model is better than the baseline method.

3.3. Results

The results obtained in this project allowed us to publish in an international conference. A paper (Oliveira, Cortez, & Areal, 2013) describing these experiments was accepted for presentation at the 3rd International Conference on Web Intelligence, Mining and Semantics (WIMS'13) in Madrid (39% acceptance rate for oral presentation) and published in the respective proceedings indexed at Scopus, DBLP and ACM Portal.

In this section, we present and discuss the results for each analyzed stock market variable, namely returns, volatility and trading volume.

3.3.1. Returns

Diverse sentiment indicators were used to infer the regression relationships between daily returns and previous day sentiment data. These sentiment indicators were created using seven lexical resources and applying two approaches of aggregation of daily indicators (i.e. S1 and S2). Tables 4 and 5 present the regression error metric values resulting from the regressions using S1 and S2 methodologies respectively. The column titled Average contains the mean of the error metric for all assets and thus is used to assess the overall value of the lexicon and sentiment approach tested. The last two columns are related with a regression model that includes two inputs: the best lexicon based variable (signaled in bold) and the baseline (r_{t-1}). Given that similar results were achieved for both R^2 and RAE, we opted to only show the R^2 metric in Tables 4 and 5. The exception is the last row of each table, which contains the RAE values.

Table 4. Returns using S1 features results (R^2 values except for last row, which includes RAE values, best R^2 value in bold)

Method	AMD	AMAZN	DELL	EBAY	GOOG	HPQ	IBM	INTC	MSFT	Average
Baseline (r_{t-1})	0.01	0.05	0.00	0.01	0.00	0.02	0.00	0.01	**0.04**	0.02
GI	0.03	0.01	0.00	0.02	**0.19**	0.00	0.11	0.16	0.00	0.06
OL	0.02	**0.16**	0.01	0.02	0.08	0.05	0.32	0.10	0.02	0.09
MSOL	**0.20**	0.05	0.02	0.04	0.17	0.01	**0.38**	0.45	0.01	0.15
MPQA	0.06	0.08	0.01	0.02	0.11	**0.08**	0.25	0.33	0.01	0.11
SWN	0.12	0.03	**0.06**	**0.04**	0.03	0.05	0.26	0.40	0.02	0.11
Emoticons	0.11	0.06	0.01	0.00	0.00	0.06	0.03	0.43	0.01	0.08
ALL	0.13	0.04	0.02	0.03	0.16	0.01	0.32	**0.46**	0.01	0.13
best $\oplus r_{t-1}$ (R^2)	0.20	0.16	0.07	0.11	0.21	0.08	0.44	0.46	0.04	0.20
best $\oplus r_{t-1}$ (RAE)	0.91	0.91	0.99	0.93	0.96	0.97	0.82	0.83	0.98	0.92

Table 5. Returns using S2 features results (R^2 values except for last row, which includes RAE values, best R^2 value in bold)

Method	AMD	AMAZN	DELL	EBAY	GOOG	HPQ	IBM	INTC	MSFT	Average
Baseline (r_{t-1})	0.01	0.05	0.00	0.01	0.00	0.02	0.00	0.01	**0.04**	0.02
GI	0.05	0.02	0.00	0.02	**0.20**	0.00	0.10	0.08	0.00	0.05
OL	0.01	**0.15**	**0.03**	0.02	0.08	0.05	0.30	0.06	0.02	0.08
MSOL	**0.24**	0.03	0.00	**0.05**	0.11	0.02	**0.34**	**0.47**	0.01	0.14
MPQA	0.05	0.08	0.00	0.02	0.07	**0.12**	0.25	0.29	0.01	0.10
SWN	0.19	0.06	0.02	0.03	0.06	0.04	0.32	0.37	0.03	0.13
Emoticons	0.11	0.06	0.01	0.00	0.00	0.06	0.03	0.43	0.01	0.08
ALL	0.14	0.06	0.00	0.03	0.13	0.04	0.28	0.38	0.01	0.12
best $\oplus r_{t-1}$ (R^2)	0.25	0.16	0.04	0.06	0.23	0.12	0.44	0.48	0.04	0.20
best $\oplus r_{t-1}$ (RAE)	0.93	0.91	0.98	0.94	0.91	0.93	0.82	0.82	0.98	0.91

Overall, the baseline method shows an almost null effect in estimating the next day returns, with R^2 values close to zero. Also, there is no added value when joining the baseline input to the best sentiment method result ($best \oplus r_{t-1}$). For few companies, such as IBM and INTC, the sentiment features seem to have an relevant contribution (e.g. R^2 values of 0.47 and 0.38) for explaining the daily returns. Nevertheless, the overall sentiment results are only slightly better than the baseline, with an average impact of 0.1 points in terms of the R^2 values for most lexicons. When comparing the sentiment methods, the results are similar for both S1 and S2 strategies. Also, few differences are found between distinct lexicons. MSOL presents the best average result for both S1 and S2. However, the overall R^2 values (0.15 and 0.14) are still low.

For demonstration purposes, Figure 6 shows the quality of the fitted results for the best model (S2 strategy and MSOL lexicon). The model is particularly good at estimating the lowest r_t value.

3.3.2. Volatility

The association between microblogging data and future volatility was assessed by linear regressions using four different specifications as previously described. Table 6 exhibits the R^2 values of these regressions, while Table 7 presents the RAE errors. Given that, in general, better results were achieved when compared with the returns regressions, we highlight the results that are better than the 0.5 threshold, for both R^2 and RAE metrics.

Here, the baseline (σ_{t-1}) is quite informative for fitting the next day volatility, with overall R^2 = 0:57 and RAE=0.50. By its own, the Twitter posting volume (n_{t-1}) does not seem useful, with an average R^2 = 0:04 and RAE=0.96. However, when combined with the baseline input ($\sigma_{t-1} \oplus n_{t-1}$ and $\sigma_{t-1} \oplus n_{t-1} \oplus n_{t-2}$), there is an increase in the fitted performance.

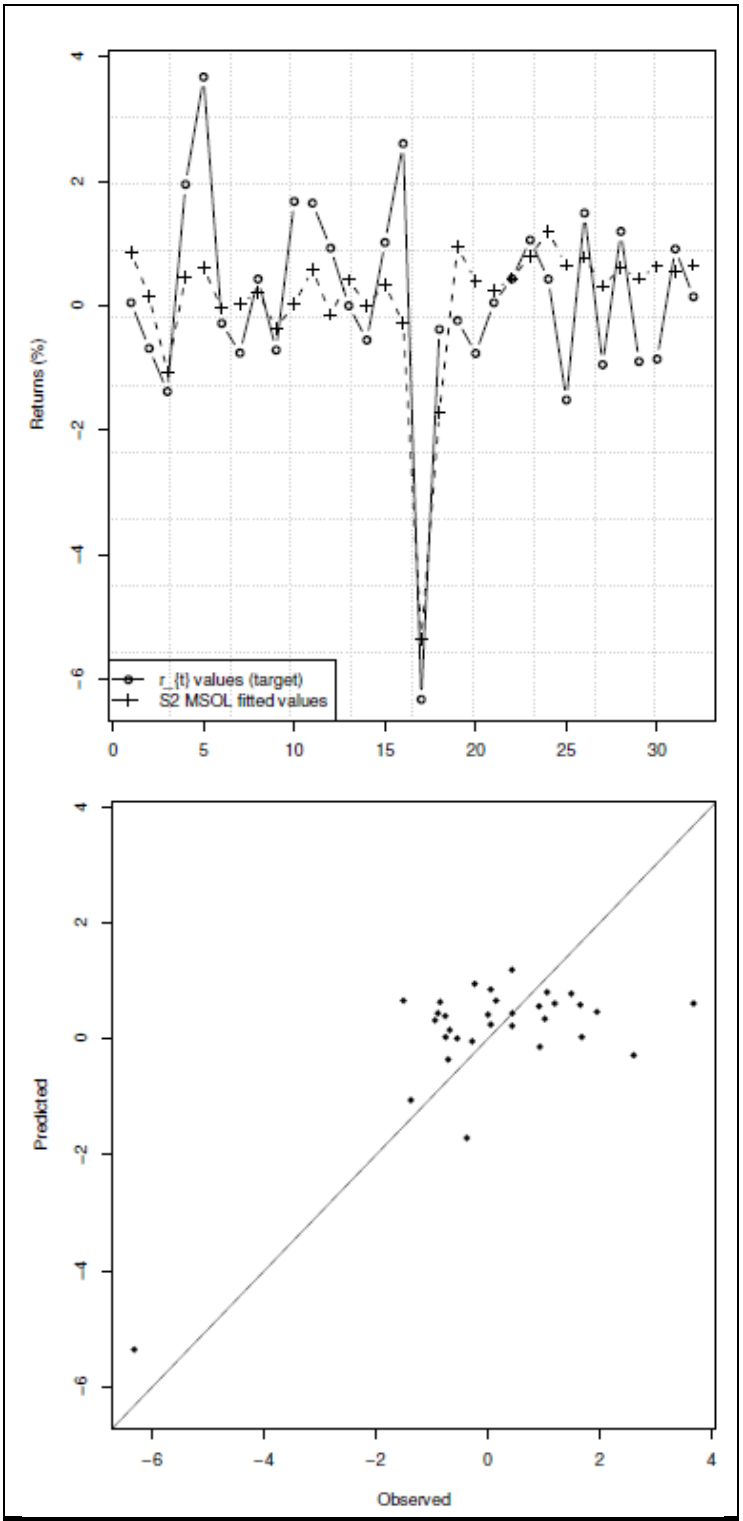


Figure 6. INTC returns and predictive values

Table 6. Volatility R^2 results (values higher than 0.5 are in bold)

Method	AMD	AMAZN	DELL	EBAY	GOOG	HPQ	IBM	INTC	MSFT	Average
Baseline (σ_{t-1})	0.32	0.36	**0.69**	**0.86**	**0.79**	0.12	**0.60**	**0.93**	0.42	**0.57**
n_{t-1}	0.00	0.05	0.00	0.03	0.09	0.01	0.14	0.00	0.07	0.04
$\sigma_{t-1} \oplus n_{t-1}$	0.36	**0.60**	**0.69**	**0.96**	**0.87**	0.12	**0.75**	**0.94**	**0.51**	**0.64**
$\sigma_{t-1} \oplus n_{t-1} \oplus n_{t-2}$	0.37	**0.71**	**0.73**	**0.97**	**0.88**	0.16	**0.77**	**0.94**	**0.53**	**0.67**

Table 7. Volatility RAE results (values lower 0.5 are in bold)

Method	AMD	AMAZN	DELL	EBAY	GOOG	HPQ	IBM	INTC	MSFT	Average
Baseline (σ_{t-1})	0.73	0.63	**0.45**	**0.20**	**0.27**	0.91	**0.44**	**0.21**	0.64	0.50
n_{t-1}	1.00	0.93	1.00	0.97	0.91	1.01	0.88	0.99	0.95	0.96
$\sigma_{t-1} \oplus n_{t-1}$	0.76	0.54	**0.45**	**0.15**	**0.28**	0.92	**0.41**	**0.19**	0.61	**0.48**
$\sigma_{t-1} \oplus n_{t-1} \oplus n_{t-2}$	0.76	**0.49**	**0.44**	**0.13**	**0.27**	0.92	**0.40**	**0.19**	0.59	**0.47**

Overall, the best results are achieved by the second combination model ($\sigma_{t-1} \oplus n_{t-1} \oplus n_{t-2}$). For this model, an average of R^2 = 0:67 and RAE=0.47 was achieved, meaning that it has a significant predictive capacity for the next day volatility. In effect, the obtained volatility fitting results are of high quality, with several results better than the threshold (AMAZN, DELL, EBAY, GOOG, IBM, INTC and MSFT).

Figure 7 shows the implied volatility and fitted values of best regression model for AMZN. We can observe that the use of the lagging values of volatility and number of tweets produces the best fit. In this particular case, the results are interesting notwithstanding the short period analyzed. As observed in the figure, the fitted model correctly identifies the raise of the highest peak (at 22 day) and the subsequent fall (at day 23).

3.3.3. Trading Volume

The results of the trading volume regressions are presented in Tables 8 (R^2 values) and 9 (RAE values). Here, the baseline (v_{t-1}) contains some predictive information, with average values of R^2 = 0:27 and RAE=0.84. More interestingly, Twitter posting volume data seems quite useful. When used by its own (n_{t-1}), the regression model outperforms in general the baseline for both error metrics (average R^2 = 0:33 and RAE=0.85). Moreover, when both inputs are combined ($v_{t-1} \oplus n_{t-1}$), the global results improve (average R^2 = 0:41 and RAE=0.81). For some companies, such as AMD, quite interesting modeling results are achieved.

In Figure 8, we present the quality of the fit for the model with best R^2 value (AMD and $v_{t-1} \oplus n_{t-1}$). The fitted values follow the diagonal line (bottom of Figure 8), suggesting an interesting fit. In particular, the raise of the highest value is correctly fitted by the model.

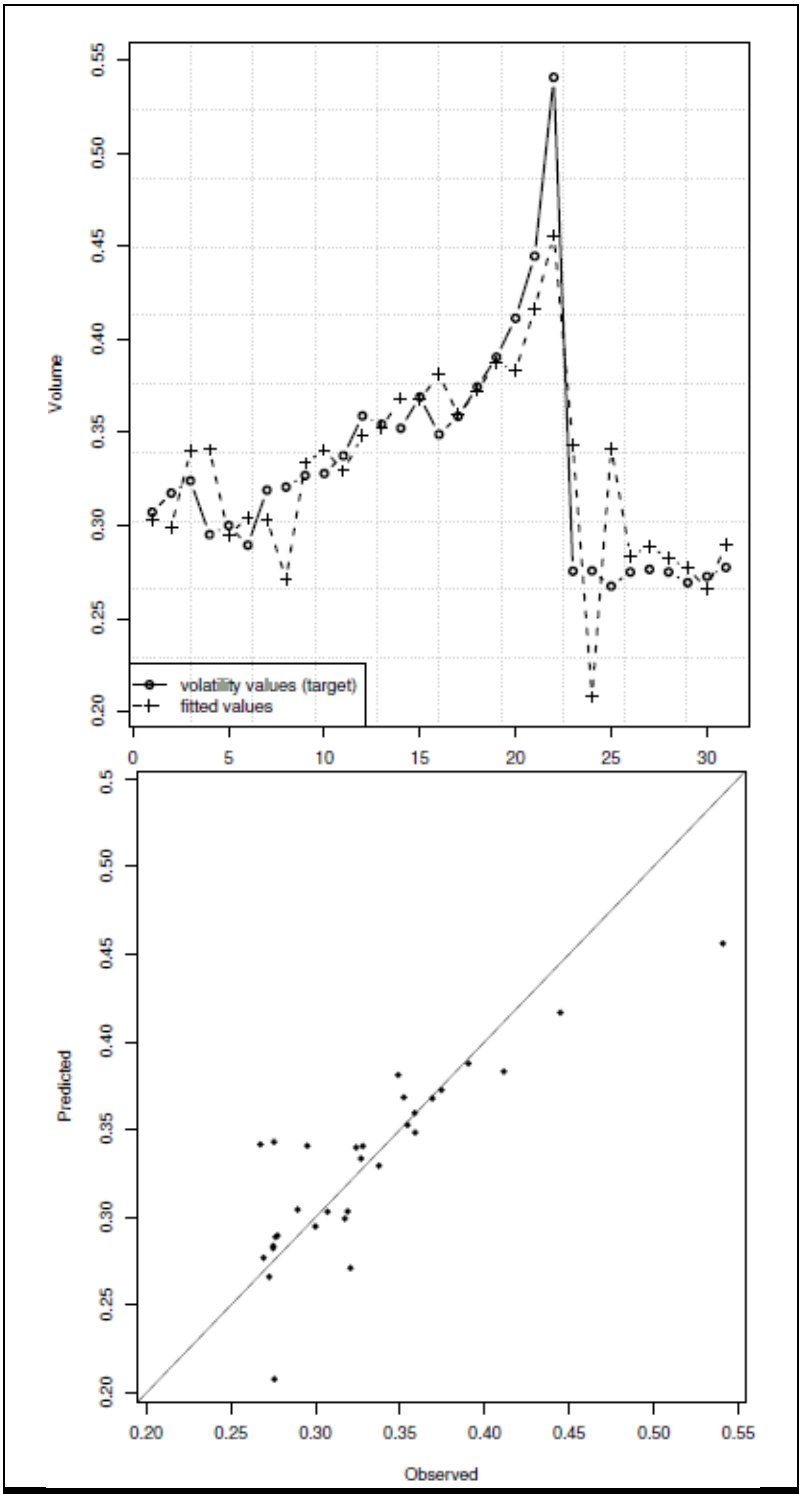


Figure 7. Volatility and fitted values for AMZN

Table 8. Volume R^2 results (best value in bold)

Method	AMD	AMAZN	DELL	EBAY	GOOG	HPQ	IBM	INTC	MSFT	Average
Baseline (v_{t-1})	0.24	0.38	0.07	0.25	0.24	0.40	0.19	0.32	0.30	0.27
n_{t-1}	0.57	**0.48**	0.12	0.39	**0.41**	0.09	0.46	0.28	0.19	0.33
$v_{t-1} \oplus n_{t-1}$	**0.58**	**0.48**	**0.14**	**0.40**	**0.41**	**0.41**	**0.54**	**0.38**	**0.32**	**0.41**

Table 9. Volume RAE results (best value in bold)

Method	AMD	AMAZN	DELL	EBAY	GOOG	HPQ	IBM	INTC	MSFT	Average
Baseline (v_{t-1})	0.88	0.86	0.90	0.85	0.85	**0.74**	0.90	0.78	**0.80**	0.84
n_{t-1}	**0.73**	**0.80**	0.87	**0.82**	**0.82**	0.92	**0.88**	0.82	0.98	0.85
$v_{t-1} \oplus n_{t-1}$	0.74	**0.80**	**0.86**	0.84	**0.82**	**0.74**	0.93	**0.76**	0.81	**0.81**

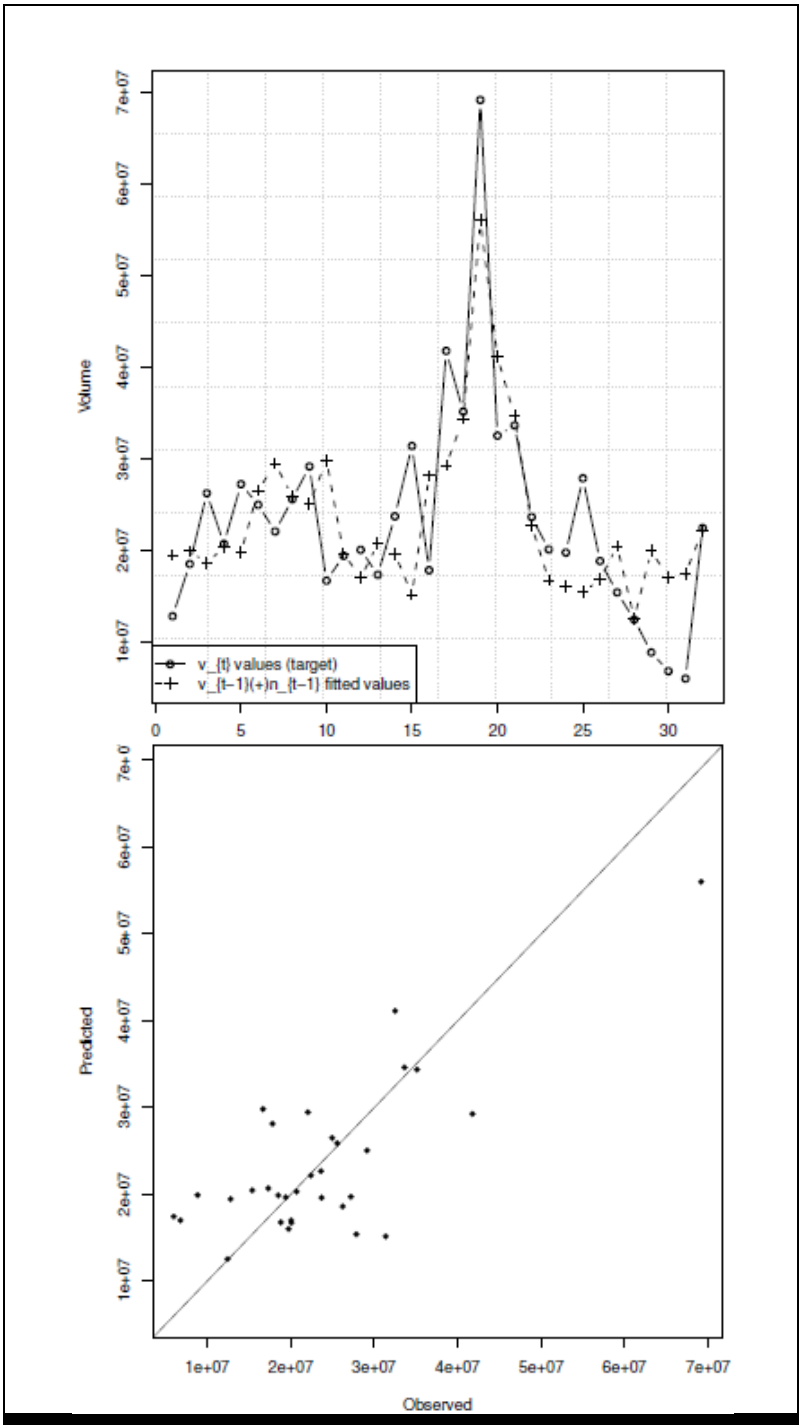


Figure 8. Trading volumes and fitted values for AMD

3.4. Discussion

The results presented here are promising, showing that information from microblog systems can be relevant for modeling the dynamics of stock market variables. We found evidence that Twitter posting volume is relevant for modeling the next day trading volume considering that most regression models based on this microblogging feature outperformed the baseline model. Moreover, the same source of data can substantially improve the modeling of volatility, provided it is used in conjunction with the previous day volatility. Posting volume seems to be an appropriate measure of investor attention in these highly posted stocks.

Confirming the scarce evidence of return predictability (Timmermann, 2008), the explored sentiment indicators did not, in general, provide significant information about the following day return. Despite some sentiment features had an important contribution (e.g., IBM, INTC), the overall sentiment results evidences scarce predictive information. Lexical resources such as MSOL, SWN and ALL enable the creation of sentiment indicators that produce average R^2 results (e.g., 0.15 and 0.13) clearly better than the baseline (0.02). This fact demonstrate that these sentiment indicators can add information to model daily returns but they are insufficient to produce reliable forecasts.

Regarding the proposed lexicons, emoticons generated poor results. Conversations about stock market utilizes less frequently the informal writing style with the presence of emoticons that is applied in common microblogging texts. The scarce presence of these sentiment symbols justifies the unsatisfactory performance of this resource. The union of the six lexical resources (ALL) performs better, achieving the second best result with the S1 approach and the third best result using the S2 approach. However, the complementarity of these lexicons is not proven in this study. The isolated utilization of the lexicon MSOL produced better results.

This project is important for this research topic because it presents important findings on some scarcely researched stock market variables.

Nevertheless, given the preliminary nature and scope of the study, and the fact that all analysis is performed in-sample, the conclusions need to be analyzed with some caution. While the results are interesting, they merit further research, for a much larger period of time with a more thorough forecasting exercise. Furthermore, the utilization of better investor sentiment indicators, produced by more sophisticated sentiment analysis algorithms, should allow a more definitive diagnosis about the predictability of returns.

4. Conclusions

4.1. Summary

The analysis of the literature about mining microblogging data to model and forecast stock market variables revealed that microblogging data may provide indicators of investor sentiment and attention in a more effective manner than traditional sources such as large scale surveys. Microblogging features demonstrated to have informative content to model and predict stock price directions (Bollen et al., 2011; Oh & Sheng, 2011), returns (Mao et al., 2011; Sprenger & Welpe, 2010), volatility (Fuehres et al., 2011) and trading volume (Sprenger & Welpe, 2010).

However, research results about the relationship between microblogging features and stock variables such as volatility, trading volume or returns, are still insufficient and inconclusive. This project reinforced the research in this topic by using Twitter features to model the next day returns, volatility and trading volume of nine major technological companies. The indicators of investor sentiment and attention created from Twitter data related to these stocks could be more representative and accurate because they have an above average posting volume. We explored five popular and large lexical resources and tested two new lexicons (emoticons and union of the remaining lexical resources) in the creation of sentiment indicators. The utilization and combination of extensive and highly tested lexical resources can contribute to produce more accurate measures of investor sentiment. Additionally, emoticons are symbols containing high emotional value and they are frequently applied to summarize the overall sentiment of the message.

The obtained results were interesting. Twitter posting volume was very useful for modeling the next day trading volume and can substantially improve the fitting of volatility when combined with the previous day

volatility. However, the sentiment indicators did not, in general, hold significant predictive information about returns.

4.2. Discussion

This project provides interesting new insights into the utilization of microblogging data to model the dynamics of stock market variables. Twitter posting volume seems to be an appropriate measure of investor attention, having relevant predictive information about trading volume and adding information to the forecasting of volatility. The isolated utilization of the number of related tweets produces goods predictions of trading volume while its combination with previous day volatility significantly improves the modeling of volatility. Therefore, the community of investors that uses Twitter to share information about stock market issues seems to be large enough to enable the creation of representative indicators, particularly in the highly posted technological sector.

The extracted sentiment indicators exhibited less capability to forecast returns, supporting the evidence that returns are difficult to predict (Timmermann, 2008).

The average results of the regression models demonstrates that the diverse sentiment indicators have generally scarce predictive information about returns. However, some sentiment features showed interesting predictive power for stocks such as IBM and INTC. Moreover, sentiment indicators produced by lexical resources such as MSOL, SWN and ALL have clearly better regressions results than the baseline model. Despite they usually produce unsatisfactory return predictions, these sentiment indicators can improve the forecasting of daily returns.

The proposed lexical resources did not present convincing results. Sentiment indicators created solely from the presence of emoticons produced poor regression results. This performance can be justified by the scarce utilization of emoticons in conversations about stock market. The merger of the six lexical resources (ALL) achieved better results. However, the utilization of the lexicon MSOL provides more quality of fit.

The obtained results are promising but some facts should be considered. This project analyzes a short period of time (32 days) and it is performed in-sample. The utilization of larger time periods and more robust forecasting procedures should produce more solid results. Moreover, there is a large margin to improve sentiment measures. The application of state of the art sentiment analysis methods may produce more accurate and complete sentiment indicators that may eventually improve the forecasting of returns.

4.3. Future Work

The implementation of this research project and analysis of the state of the art related with mining microblogging data to model and forecast stock market behavior allowed us to identify some research opportunities, such as:

- analyze a significant period of time with more thorough forecasting exercises in order to perform a more robust assessment about the relevance of microblogging data for forecasting stock market behavior.
- apply state of the art sentiment analysis algorithms able to accurately label the sentiment of the entire message but also to extract the opinion about some relevant stock features (e.g., financial parameters, product performance). Moreover, this algorithms should address the specificities of social media contents (e.g., profusion of grammatical errors, specific structure and terminology) and stock market vocabulary. These methods could produce more accurate and complete sentiment indicators that can potentially improve the forecasting of stock market behavior.
- perform social network analysis to identify influential users, to assess the reputation of the authors and to understand how opinion is formed.
- combine diverse sources of web data such as social media services and internet searches. These sources have distinct

characteristics that can be complementary and enable better predictions. Google searches represent a superior number of users and is a direct measure of attention regarding a comprehensive set of topics. Blogs have more complete opinionated content because their authors can freely describe their personal views and share important information. Microblogging contents have greater objectivity, interactivity and posting frequency. A deeper analysis of the relevance of different sources of Web data and their dynamic combination can result in better forecasting ability of financial indicators.

- explore inferior posting periodicities. The daily frequency has been the minimum periodicity applied in this research topic. However, it does not allow near real-time sentiment report that could be used throughout the trading day. The utilization of a timely analysis, performed quickly after posting (e.g. minutes or hours) may enable a better exploitation of sentiment effects in stock market.

References

Aggarwal, C. C., & Zhai, C. (2012a). An Introduction to Text Mining. In C. C. Aggarwal & C. Zhai (Eds.), *Mining Text Data* (pp. 1–10). Springer. doi:10.1007/978-1-4614-3223-4

Aggarwal, C. C., & Zhai, C. (2012b). A Survey of Text Classification Algorithms. In C. C. Aggarwal & C. Zhai (Eds.), *Mining Text Data* (pp. 163–222). Springer. doi:10.1007/978-1-4614-3223-4

Aggarwal, C. C., & Zhai, C. (2012c). A Survey of Text Clustering Algorithms. In C. C. Aggarwal & C. Zhai (Eds.), *Mining Text Data* (pp. 77–128). Springer. doi:10.1007/978-1-4614-3223-4

Amihud, Y., Mendelson, H., & Pedersen, L. H. (2005). Liquidity and Asset Prices. *Foundations and Trends® in Finance*, *1*(4), 269–364. doi:10.1561/0500000003

Antweiler, W., & Frank, M. (2004). Is All That Talk Just Noise? The Information Content of Interest Stock Message Boards. *Journal of Finance*, *59*(3), 1259.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In N. C. C. Chair, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, … D. Tapias (Eds.), *Proceedings of the Seventh conference on International Language Resources and Evaluation LREC10* (Vol. 0, pp. 2200–2204). European Language Resources Association (ELRA). Retrieved from http://www.lrec-conf.org/proceedings/lrec2010/pdf/769_Paper.pdf

Baker, L. D., & McCallum, A. K. (1998). Distributional clustering of words for text classification. (W. B. Croft, A. Moffat, C. J. Van Rijsbergen, R. Wilkinson, & J. Zobel, Eds.)*Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval SIGIR 98*, *104*(1), 96–103. doi:10.1145/290941.290970

Bank, M., Larch, M., & Peter, G. (2011). Google search volume and its influence on liquidity and returns of German stocks. *Financial Markets and Portfolio Management*, *25*(3), 239–264.

Barber, B. M., & Odean, T. (2008). All That Glitters: The Effect of Attention and News on the Buying Behavior of Individual and Institutional Investors. *Review of Financial Studies*, *21*(2), 785–818.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. (M. Jordan, J. Kleinberg, & B. Schölkopf, Eds.)*Pattern Recognition* (Vol. 4, p. 738). Springer. doi:10.1117/1.2819119

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1–8.

Brody, S. (2010). An Unsupervised Aspect-Sentiment Model for Online Reviews. *Computational Linguistics*, (June), 804–812. Retrieved from http://www.aclweb.org/anthology/N10-1122

Covington, M. A. (2001). A Fundamental Algorithm for Dependency Parsing. *Machinery*, 95–102. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.136.7335&rep=rep1&type=pdf

Da, Z., Engelberg, J., & Gao, P. (2011). In Search of Attention. *Journal of Finance*, *LXVI*(5), 1461–1499.

Daniel, K., Hirshleifer, D., & Teoh, S. H. (2002). Investor psychology in capital markets: evidence and policy implications. *Journal of Monetary Economics*, *49*(1), 139–209.

Das, S. R., & Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web. *Management Science*, *53*(9), 1375–1388.

De Choudhury, M., Sundaram, H., John, A., & Seligmann, D. D. (2008). Can blog communication dynamics be correlated with stock market activity? *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia HT 08*, 55–60. doi:10.1145/1379092.1379106

Dimpfl, T., & Jank, S. (2011). Can internet search queries help to predict stock market volatility? *German Research*.

Ding, R., & Hou, W. (2011). Retail Investor Attention and Stock Liquidity. *SSRN*. Retrieved from http://ssrn.com/abstract=1786762

Fayyad, U., Piatetsky-shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. (U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy, Eds.)*AI Magazine*, *17*(3), 37–54. doi:10.1609/aimag.v17i3.1230

Feldman, R., & Sanger, J. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. (J. Sanger, Ed.)*Imagine* (Vol. 34, p. 410). Cambridge University Press. doi:10.1179/1465312512Z.00000000017

Freitag, D., & McCallum, A. (2000). Information Extraction with HMM Structures Learned by Stochastic Optimization. In *AAAIIAAI* (Vol. 3, pp. 584–589). Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999. doi:10.1145/1105664.1105679

Fuehres, H., Zhang, X., & Gloor, P. A. (2011). Predicting Stock Market Indicators Through Twitter “I hope it is not as bad as I fear.” *Procedia - Social and Behavioral Sciences*. doi:10.1016/j.sbspro.2011.10.562

Ganapathibhotla, M., & Liu, B. (2008). Mining opinions in comparative sentences. *Camera*, *1*(August), 241–248. doi:10.3115/1599081.1599112

Gu, B., Konana, P., Liu, A., Rajagopalan, B., & Ghosh, J. (2006). Identifying Information in Stock Message Boards and Its Implications for Stock Market Efficiency. In *Workshop on Information Systems and Economics 2006*.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. (Springer, Ed.)*The Mathematical Intelligencer* (Vol. 27, p. 745). Springer. doi:10.1177/001112877201800405

Hirshleifer, D., & Teoh, S. H. (2003). Limited attention, information disclosure, and financial reporting. *Journal of Accounting and Economics*, *36*, 337–386.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining KDD 04*, *04*(2), 168. doi:10.1145/1014052.1014073

Hu, X., & Liu, H. (2012). Text Analytics in Social Media. In C. C. Aggarwal & C. Zhai (Eds.), *Mining Text Data* (pp. 385–414). Springer. doi:10.1007/978-1-4614-3223-4

Jakob, N. (2010). Extracting Opinion Targets in a Single- and Cross-Domain Setting with Conditional Random Fields. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, (October), 1035–1045. Retrieved from http://www.aclweb.org/anthology/D10-1101

Jiang, G. J. (2005). The Model-Free Implied Volatility and Its Information Content. *Review of Financial Studies*, *18*(4), 1305–1342. doi:10.1093/rfs/hhi027

Jiang, J. (2012). Information Extraction from Text. In C. C. Aggarwal & C. Zhai (Eds.), *Mining Text Data* (pp. 11–42). Springer. doi:10.1007/978-1-4614-3223-4

Jin, W., & Ho, H. H. (2009). A Novel Lexicalized HMM-based Learning Framework for Web Opinion Mining. (L. Bottou & M. Littman, Eds.)*Techniques*, 1–8. doi:10.1145/1553374.1553435

Jindal, N., & Liu, B. (2006). Mining Comparative Sentences and Relations. *Artificial Intelligence*, *21*(2), 1331–1336. doi:10.1107/S0108270189000326

Joseph, K., Wintoki, M. B., & Zhang, Z. (2011). Forecasting abnormal stock returns and trading volume using investor sentiment: Evidence from online search. *International Journal of Forecasting*, *27*(4), 1116–1127.

Jurafsky, D., & Martin, J. H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. (S. Russel & P. Norving, Eds.)*Computational Linguistics* (Vol. 163, p. 934). Prentice Hall. doi:10.1162/089120100750105975

Kharratzadeh, M., & Coates, M. (2011). Weblog Analysis for Predicting Correlations in Stock Price Evolutions. *International AAAI Conference on Social Media and Weblogs*, 491–494.

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In C. E. Brodley & A. P. Danyluk (Eds.), *MACHINE LEARNINGINTERNATIONAL WORKSHOP THEN CONFERENCE* (Vol. pages, pp. 282–289). Citeseer. doi:10.1038/nprot.2006.61

Leary, D. E. O. (2011). Blog mining-review and extensions : "From each according to his opinion." *Decision Support Systems*.

Lewis, D. D., N'{e}dellec, C., & Rouveirol, C. (1998). Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval. *Machine Learning: ECML-98*, 4–15. doi:10.1007/BFb0026666

Lin, C., & He, Y. (2009). Joint sentiment/topic model for sentiment analysis. *peoplekmiopenacuk*, 375–384. Retrieved from http://oro.open.ac.uk/23786/

Liu, B., & Zhang, L. (2012). A Survey of Opinion Mining and Sentiment Analysis. In C. C. Aggarwal & C. Zhai (Eds.), *Mining Text Data* (pp. 415–464). Springer. doi:10.1007/978-1-4614-3223-4

Mao, H., Counts, S., & Bollen, J. (2011). Predicting Financial Markets: Comparing Survey, News, Twitter and Search Engine Data. *arXiv.org*.

McCallum, A., & Nigam, K. (1998). *A comparison of event models for naive bayes text classification*. *Workshop on Learning for Text Categorization* (pp. 41–48). doi:10.1.1.46.1529

Merton, R. C. (1987). A Simple Model of Capital Market Equilibrium with Incomplete Information. *Journal of Finance*, *42*(3), 483–510.

Mohammad, S., Dunne, C., & Dorr, B. (2009). 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* (Vol. 2, pp. 599–608). Association for Computational Linguistics. doi:10.3115/1699571.1699591

Nenkova, A., & McKeown, K. (2012). A Survey of Text Summarization Techniques. In C. C. Aggarwal & C. Zhai (Eds.), *Mining Text Data* (pp. 43–76). Springer. doi:10.1007/978-1-4614-3223-4

Nofsinger, J. R. (2005). Social Mood and Financial Economics Social Mood and Financial Economics. *Journal of Behavioral Finance*, *6*(3), 144–160.

Oh, C., & Sheng, O. R. L. (2011). Investigating Predictive Power of Stock Micro Blog Sentiment in Forecasting Future Stock Price Directional Movement. In *ICIS 2011 Proceedings*. Shanghai, China.

Oliveira, N., Cortez, P., & Areal, N. (2013). 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 (WIMS '13)*. Madrid, Spain: ACM. doi:10.1145/2479787.2479811

Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. (C. C. Aggarwal & C. Zhai, Eds.)*Foundations and Trends in Information Retrieval*, *2*(2), 1–135.

Peterson, R. L. (2007). Affect and Financial Decision-Making: How Neuroscience Can Inform Market Participants. *Journal of Behavioral Finance*, *8*(2), 70–78.

Porter, M. F. (1980). An algorithm for suffix stripping. (K. S. Jones & P. Willet, Eds.)*Program*, *14*(3), 130–137. doi:10.1108/00330330610681286

Ruiz, E. J., Hristidis, V., Castillo, C., Gionis, A., & Alejandro, J. (2012). Correlating Financial Time Series with Micro-Blogging Activity. In

Proceedings of the fifth ACM international conference on Web search and data mining (pp. 513–521). New York, USA: ACM.

Sabherwal, S., Sarkar, S. K., & Zhang, Y. (2008). Online talk: does it matter? *Managerial Finance*, *34*(6), 423–436.

Schütze, H., & Silverstein, C. (1997). Projections for Efficient Document Clustering. In *Proc SIGIR* (Vol. 31, pp. 74–81). ACM Press. doi:10.1145/278459.258539

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, *14*(3), 199–222. doi:10.1023/B:STCO.0000035301.49549.88

Sprenger, T. O., & Welpe, I. M. (2010). Tweets and Trades: The Information Content of Stock Microblogs. *Social Science Research Network Working Paper Series*, 1–89.

Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. (M. I. T. Press, Ed.)*The MIT Press* (Vol. 08, p. 651). MIT Press. Retrieved from http://www.webuse.umd.edu:9090/

Su, Q., Xu, X., Guo, H., Guo, Z., Wu, X., Zhang, X., & Swen, B. (2008). Hidden Sentiment Association in Chinese Web Opinion Mining. *Distribution*, *98*(6 Pt 1), 959–968. doi:10.1145/1367497.1367627

Team, R. C. (2012). R: A language and environment for statistical computing. Retrieved from http://www.r-project.org

Timmermann, A. (2008). Elusive return predictability. *International Journal of Forecasting*, *24*(1), 1–18. doi:10.1016/j.ijforecast.2007.07.008

Titov, I., & McDonald, R. (2008). Modeling Online Reviews with Multi-grain Topic Models. *Proceeding of the 17th international conference on World Wide Web WWW 08*, *3*(135), 111. Retrieved from http://arxiv.org/abs/0801.1063

Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. (M. Hearst & M.

Ostendorf, Eds.)*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. doi:10.3115/1073445.1073478

Turban, E., Sharda, R., Aronson, J. E., & King, D. (2007). *Business Intelligence: A Managerial Approach* (p. 264). Prentice Hal.

Vlastakis, N., & Markellos, R. N. (2012). Information Demand and Stock Market Volatility. *Journal of Banking & Finance*, *36*(6), 1808–1821.

Wilson, T., Wiebe, J., & Hoffmann, P. (2009). Recognizing Contextual Polarity: An Exploration of Features for Phrase-Level Sentiment Analysis. *Computational Linguistics*, *35*(3), 399–433. doi:10.1162/coli.08-012-R1-06-90

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques* (2nd ed.). Morgan Kaufmann.

Wysocki, P. D. (1999). Cheap Talk on the Web: The Determinants of Postings on Stock Message Boards. *Changes*, *i*.

Yu, Y., Duan, W., & Cao, Q. (2013). The impact of social and conventional media on firm equity value: A sentiment analysis approach. *Decision Support Systems*, *55*(4), 919–926.

Zaima, A., & Kashner, J. (2003). Data Mining Primer for the Data Warehouse Professional. *Intelligence*, 44–54.

Zhang, Y. (2009). Determinants of Poster Reputation on Internet Stock Message Boards. *American Journal of Economics and Business Administration*, *1*(2), 114–121.