POLITECNICO DI MILANO

Corso di Laurea Magistrale in Ingegneria Informatica

Dipartimento di Elettronica e Informazione

Design and validation of a forecasting trading system based on Twitter

Supervisor: Prof. Chiara Francalanci

Assistant Supervisor: Ing. Alessandro Poli

Master Graduation Thesis by:

Alberto Maggioni, Student ID 765576

Luca Mazzoni, Student ID 765813

Academic Year 2011-2012

Table of contents

Table of contents I

List of figures III

List of tables V

Abstract VI

Sommario VII

Acknowledgments VIII

1 Introduction 1

2 State of the art 4

2.1 Social Media 4

2.1.1 Introduction 4

2.1.2 Social networks population and usage 7

2.2 A sociological perspective 10

2.2.1 Social networks structure 10

2.2.2 Social networks diffusion models 12

2.2.3 Real-world applications 14

2.3 Automated comprehension of text 15

2.3.1 Background on Natural Language Processing 15

2.3.2 Sentiment analysis 18

2.3.3 Applications 21

2.3.4 Sentiment analysis tools 22

2.4 Prediction from social media 23

2.4.1 Motivation 23

2.4.2 The early Web 2.0 days 24

2.4.3 From blogs to social media platforms 26

2.4.4 Exploiting the sentiment 29

2.5 An introduction on algorithmic trading 31

2.6 Using the sentiment to predict the stock market 37

3 Qualitative and exploratory analyses 49

3.1 Why Twitter? 50

3.2 Twitter as a news medium 51

3.3 Linear and nonlinear models 54

3.4 Initial experiments 56

4 Methodology 60

4.1 Modelling the social mood 60

4.1.1 Model definition 61

4.1.2 Implementation 70

4.2 Predicting the stock market 76

4.2.1 The predictive algorithm 77

4.2.2 Complexity 92

5 Experimental study and analysis of the results 95

5.1 Which sector? 96

5.2 Which brand? 97

5.3 Setting-up the crawler 101

5.4 Brand model refinement 101

5.5 Testing the sentiment analysis component 104

5.6 Insights on returns and on Twitter posts 106

5.7 Assumptions testing on linear regression models 116

5.8 Setting up the algorithm 120

5.9 Simulation Results 123

5.9.1 Statistical properties of daily P&L 126

5.9.2 Where do predictions come from? 128

6 Conclusions and future developments 130

Appendix 133

Bibliography 136

List of figures

Figure 3.1: Overlap between the curve of the prices and the histogram of the sentiment during March 20, 2012 58
Figure 4.1: Activity diagram of the brand model definition and sentiment analysis 64
Figure 4.2: The financial categories hierarchy 69
Figure 4.3: The overall conceptual architecture 78
Figure 4.4: Activity diagram of the core part of the algorithm 80
Figure 4.5: Scheme that depicts the prediction series resulting from the k times reiteration of the prediction part 89
Figure 4.6: The comparison between different prediction series 92
Figure 4.7: Timeline that represents the two posts groups that are gathered from Twitter assuming that the chosen stock market opens at 9:30 and closes at 16:00 93
Figure 5.1: Monthly posts volume 107
Figure 5.2: Description of the graphical elements of a box plot 108
Figure 5.3: Box plot of non-financial posts 109
Figure 5.4: Box plot of financial posts 109
Figure 5.5: Box plot of total positive posts 110
Figure 5.6: Box plot of total negative posts 110
Figure 5.7: Box plot of all posts per session 111
Figure 5.8: Box plot of neutral posts per session 112
Figure 5.9: Box plot of positive financial posts per session 113
Figure 5.10: Box plot of DJTATO returns per session 115
Figure 5.11: Frequencies of DJTATO returns 115
Figure 5.12: Number of single unique models per TFW 117
Figure 5.13: Number of single models that produce more than 10% of the predictions 118

Figure 5.14: Fraction of the whole predictions yielded by each single class of model 118
Figure 5.15: Average per cent of the total predictions per class of models 119
Figure 5.16: Empirical approach scheme 120
Figure 5.17: Investment results 124
Figure 5.18: The box plots of the profits and losses of the simulation (on the left) and of the daily DJTATO returns (on the right) 127
Figure 5.19: The histograms of the DJTATO returns (on the left) and of the profit and losses obtained by the simulation (on the right) 127

List of tables

Table 2.1: Classification of social media 6
Table 2.2: Facebook and Twitter age distribution 7
Table 3.1: Posting frequency of the most active users on Twitter 53
Table 3.2: Frequencies of appearances of news sites on Google 53
Table 5.1: DJTATO components and their relative weight 98
Table 5.2: Manual categorization results 105
Table 5.3: Manual sentiment evaluation results 105
Table 5.4: Precision and recall levels of the categorization 106
Table 5.5: Precision and recall levels of the sentiment evaluation 106
Table 5.6: DJTATO returns 113
Table 5.7: Some statistical measures of the DJTATO returns 116
Table 5.8: Values obtained for parameters 123
Table 5.9: Global returns of the chosen investments 124
Table 5.10: Some investment indicators 126

Abstract

In the last years, the exploitation of the *sentiment* expressed by the users on the various social networks for analysing the reputation of a given *brand* (known as *Web Reputation*) has reached a high level of maturity and many companies started their business to sell this kind of service to bigger firms. Even more recently there have been proposed some works that exploit the *sentiment* in order to predict real world outcomes, such as the variations of stock market indices.

This work proposes an innovative approach to the prediction of financial instrument prices, which combines a semantic *sentiment analysis* technique together with an algorithm specifically designed to take advantage of both the *sentiment* and the past values of the chosen financial instrument, in order to ensure the maximization of the obtainable profits by trading such product on a given market.

We will start by presenting a model of the domain of interest, that drives the *sentiment analysis* engine with the purpose of precisely identify the subject to which the *sentiment* refers to. Then we will present the algorithm that we have designed in order to exploit that knowledge in order to maximize the investment return.

The distinctive characteristic of this work is indeed represented by the combination of those two elements that, as shown by an investment simulation, allowed us to obtain a much better performance with respect to that of the selected financial instrument.

Sommario

Negli ultimi anni si è consolidato l'utilizzo del *sentiment* espresso dagli utenti sui vari social network per l'analisi della reputazione di una data azienda (la cosiddetta *Web Reputation*) e sono nate molte società con l'obiettivo di vendere questo servizio ad aziende di grandi dimensioni. Ancor più recentemente sono stati proposti alcuni lavori che sfruttano il *sentiment* per predire fenomeni del mondo reale, come ad esempio le variazioni di alcuni indici di mercati azionari.

Questo lavoro propone un approccio innovativo alla predizione dei prezzi di strumenti finanziari, che combina una tecnica semantica di *sentiment analysis* con un algoritmo progettato appositamente per trarre vantaggio dall'utilizzo del *sentiment* e dei valori passati dello strumento finanziario scelto, in modo da garantire la massimizzazione dei profitti ottenibili scambiando sul mercato tale prodotto.

Sarà dapprima proposto un modello del dominio d'interesse, utilizzato dal motore di *sentiment analysis* con lo scopo di identificare con precisione il soggetto al quale il *sentiment* si riferisce. In seguito verrà presentato l'algoritmo progettato per sfruttare tale conoscenza al fine di massimizzare il rendimento dell'investimento.

Il carattere distintivo del presente lavoro è rappresentato proprio dalla combinazione di questi due elementi che, come mostrato mediante una simulazione d'investimento, hanno permesso di ottenere una performance migliore rispetto a quella dello strumento finanziario scelto.

Acknowledgments

A special thanks to Chiara Francalanci, Alessandro Poli and Francesco Merlo for their kind support and the contagious enthusiasm with which they followed our work.

Luca and Alberto

To my father, for his esteem that is greater than I deserve.
To my mother, whose life stands as a proof of what can be achieved with willpower and determination.
To my sister, who has been a source of inspiration since I was a child.
To my cat, which always remembers me that the happiness recipe is simple: eat and sleep, that's all!

To my girlfriend, who is like my Ferrari: I don't have one either! (after having read thousands of tweets a citation is needed)
To my schoolmates, for having shared the joys and pains of life at Politecnico.
To my friends, for not having insulted me when I was forced to study on Saturday nights.
To Cata, a friend that I will always keep in my heart.

To myself, for being what I am and not being what I am not.

To all those who smiled while reading these lines and now think that I am quite crazy: you are probably right, as I have conceived these acknowledgements in my bed, instead of trying to sleep.

Alberto

A sincere and deep gratitude to my father Roberto who taught me to never give up and to my mother Raffaella for her kind support during this years at Politecnico.

To my sister for having challenged my teaching skills.

To the rest of my family and especially to my grandmother Antonia for her prayer and for her devotion to the family.

To my lifelong friends Alessandro and Nicola for the great but unfortunately rare moments together.

To my friends and colleagues at Politecnico for having shared this unique life experience.

To Rossella and for our life together: the best is yet to come.

Luca

Chapter 1

Introduction

On the web and, particularly, on social media, people express their opinions and provide ratings of a number of products and services. This thesis starts from the assumption that people's opinions and ratings have an impact on the financial performance of the companies providing those products and services. Although this impact is more direct on companies' revenues, we believe that there might be a consequent effect also on the stock prices of the subset of companies that are traded in the stock exchange. Since there exists a delay between opinions and financial performance our goal is to exploit users' buzz to predict the financial market. Although behavioural finance explains that there exist long-term investment waves that are set by large investors rather than people and their mood, we wish to verify the idea that smaller investors make shorter term investment decisions based on the general mood on an industry or on a specific company.

The main contribution of this thesis is to design a predictive trading algorithm that takes people's sentiment as an input and returns a short-term investment decision. Therefore, this thesis is positioned in the algorithmic trading field, where specific algorithms are designed in order to perform trading activities in place of humans.

We opted for taking Twitter as the reference social platform for its immediacy and we decided to focus our attention on the automotive sector as it is well chatted on the chosen medium.
The first step of this work has been the development of a *brand model* in the attempt to enrich a *sentiment analysis* tool with a *domain knowledge* specifically targeted on automotive firms and their financial aspects. This requires the definition of a set of *keywords*, *categories* and *semantic concepts*.
Then we have designed an adaptive predictive algorithm that keeps track of past performance measures and that tries to constantly change its behaviour in order to better model both *sentiment* and financial data.
Finally, the algorithm has been run on a training period in order to tune some fundamental parameters and tested on a subsequent period by simulating an investment on the selected financial instrument, which enabled us to gain 14.7% of the initial capital, with respect to a return of just 2.8% achieved by the same instrument.

The thesis is structured as follow.

Chapter 2 inspects the state of the art related to many different topics with the purpose of illustrating the basis over which our work has been built. First, it establishes some definitions about the social media world accompanied by some statistics on their diffusion and common usages. A sociological perspective is also offered in order to explain their internal dynamics and the resulting applications. Second, it explains in detail the most common *natural language processing* and *sentiment analysis* approaches, describing the pros and cons of the different techniques and the available tools. Third, a brief glance to the algorithmic trading field is provided in order to illustrate some widely applied trading strategies. Finally, recent works aimed to predict generic real world outcomes and stock market trends are presented.

Chapter 3 describes the preliminary work that was conducted in order to better understand the problem and to trace the way to be followed for its solution.

Chapter 4 explains the proposed methodology by listing all the steps needed to enrich the *sentiment analysis* tool, the assumptions made and the functioning of the designed predictive algorithm.

Chapter 5 reports detailed explanations of the choices made in order to apply the proposed methodology and some quality measures that assess the adequacy of both the *brand model* and the *sentiment analysis* component. It also provides some data in order to prove all the assumptions made during the development of the algorithm. Moreover, financial and *sentiment* data are analysed from a statistical perspective. Lastly, it shows the results of an investment simulation whose performance is measured and compared to a benchmark in order to evaluate the effectiveness of the entire approach.

Chapter 6 summarize the entire work and the obtained results and provides some suggestions for future developments.

Finally, the Appendix contains the list of *semantic concepts* for each *category*.

Chapter 2

State of the art

2.1 Social Media

2.1.1 Introduction

> As of January 2009, the online social networking application Facebook registered more than 175 million active users. To put that number in perspective, this is only slightly less than the population of Brazil (190 million) and over twice the population of Germany (80 million)! At the same time, every minute, 10 hours of content were uploaded to the video sharing platform YouTube. And, the image hosting site Flickr provided access to over 3 billion photographs, making the world-famous Louvre Museum's collection of 300,000 objects seem tiny in comparison. [1]

These huge numbers, even if quite dated, clearly demonstrate the importance of what has been defined "the social media revolution" [2]. But let's first clarify the concept of social media, which should be distinguished from that of *Web 2.0* and *User Generated Content* (*UGC*). According to Kaplan and Haenlein [1], *Web 2.0*, that is a new way of utilizing the World Wide Web as a platform whereby content and applications are continuously modified by all users, is the ideological and

technological foundation for the evolution of social media. Conversely, *UGC* is a term applied to describe the various forms of media content that must accomplish some simple requirements [3]:

1. It must be accessible to all the Internet users or at least to some of the users of a social networking site (so e-mails or instant messages are not examples of *UGC*).
2. It must be creative (a mere copy of an existing article without any commenting is not *UGC*).
3. It should not have been created for commercial purposes.

Therefore, they defined social media as "a group of Internet-based applications that are built on the ideological and technological foundations of *Web 2.0*, and that allow the creation and exchange of *User Generated Content*". This broad definition includes a lot of different websites, thus they proposed a further classification that relies on the following notions:

- *Social presence:* it is the achievable physical contact between two communication partners; the social influence that the communication partners have on each other's behaviour increases with an increase in social presence.
- *Media richness:* it refers to the rate at which information flows and so the effectiveness in resolving ambiguity and uncertainty.
- *Self-presentation:* it is the everybody's desire to shape other people's impressions of himself through the creation of a personal webpage; different media allow different types of self-presentation;
- *Self-disclosure:* it refers to how much personal information a user is required to give on a given medium.

These orthogonal dimensions of analysis lead to the classification of social media shown in Table 2.1. Text-based applications, such as collaborative projects, score lowest with respect to social presence and media richness,

while on the highest level we can find social worlds, that try to replicate real life scenarios and require a great level of self-disclosure.

Table 2.1: Classification of social media

		Social presence / Media richness		
		Low	Medium	High
Self-presentation / Self-disclosure	High	Blogs	Social networking sites (e.g. Facebook)	Virtual social worlds (e.g. Second Life)
	Low	Collaborative projects (e.g. Wikipedia)	Content communities (e.g. YouTube)	Virtual game worlds (e.g. World of Warcraft)

We will focus on social networking sites (and in particular on Twitter) that are applications through which people can share personal information (in the form of photos, videos, audio files or plain text) and communicate with other users. Facebook and Twitter are the top two social networking sites in terms of overall number of users [4]. This type of social media is also the most widely used. “Online social networks are changing the way people communicate, work and play, and mostly for the better” said Martin Giles, a technology correspondent for the journal “The Economist” [5]. There are plenty of statistics that shows their exponential growth: Americans spend more time on social networks than on email [6], there are more Facebook users than cars [7], 1 million accounts are currently added to Twitter every day [8] and so on.

2.1.2 Social networks population and usage

As said before, the two biggest social networks in terms of number of users are Facebook and Twitter. Therefore, we report here some statistics about the demographics of these two websites, which are hoped to provide valuable insights. They are taken from [9] and crosschecked on [10]. Facebook has 845 millions of active users, of which 43% males and 57% females. Twitter, with its 127 millions of active users, presents quite the same sex ratio: 43% males and 57% females. They are also similar with respect to the age distribution, as shown in Table 2.2.

Table 2.2: Facebook and Twitter age distribution

Age	Facebook	Twitter
0-24	14%	19%
25-34	18%	23%
35-44	22%	25%
45+	46%	33%

The same hold for the average level of instruction, with the great majority of users with education at or beyond the "some college" level.

Now, let's go through the list of common practices in social networks usage. Starting from the basis, that is, from the first examples of *UGC*, Schiano et al. [11] reported some major motivations for blogging: reporting facts from one's life, sharing comments and opinions, expressing deeply felt emotions, providing information about some topics and building communities. We obviously expect these to be the same reasons that drive social networks users, and this is the case.

Following Java et al. [12], that analysed the typical behaviours on the early Twitter, we can distinguish three main types of user intention:

information sharing, information seeking, and friendship-wise relationship. The following are the corresponding user types:

- *Information Source:* a user with a large number of *followers* who post updates on regular intervals or infrequently (but with valuable content). This category also includes automated tools that post news on Twitter.
- *Friends:* many users have friends, family and co-workers on their friends or *followers* lists and share information with them.
- *Information Seeker:* a person with a limited posting activity who aims to gather information by following other users.

It must be pointed out that a user may fall in more than just one category, playing different roles in different communities. Finally, they classified the most typical posts as follows:

- *Daily Chatter:* posts regarding daily habits or people's on-going actions. It is the largest and most common kind of tweet (as confirmed by a 2009's study of Peeranalytics [13]).
- *Conversations:* tweets that contain the @ symbol followed by a username in order to send a direct message to a friend or comment a previous post.
- *Sharing information/URLs:* posts that contain some URL in them.
- *Reporting news:* many real or automated users share news and their opinions about current events.

By the previous categorization, it follows that whatever analysis of social media content should take into account the need of separating the so-called "pointless babble" from other more interesting kinds of chatter. In addition to this research paper, conversation has been found to be a widely spread and increasing purpose in social networks usage also by [14]. They pointed out that this type of communication contributes

towards varying the otherwise rather monothematic content of tweets (the original purpose of Twitter was to answer the question "What are you doing?" [15]), with the introduction of new forms of interactions, such as collaborative problem solving. A further poll by Lee Odden [16] reported interesting results: for instance, 7% of the users declared to use Twitter specifically to promote specific content. However, the sample was a bit polarized and limited to the blog visitors. Kwak et al. [17] emphasized the importance of Twitter as a news medium (as a matter of fact the official question has been changed into "What's happening?" [15]) and as an ideas spreader; in fact, the mechanism of *retweeting* allows tweets to reach a huge audience, even if they were posted by people with a limited amount of *followers*.

Finally, let's take a brief look at how enterprises act on social networks. Jansen et al. [18] showed that the word of mouth, i.e. the process of conveying information from person to person that influences customer buying decisions, plays an important role in microblogging services. Therefore, a number of businesses and organizations started using Twitter or similar to keep in touch with their stakeholders, to support the creation of brand communities and for viral marketing campaigns (cfr. [19] for a partial list of brands employing Twitter). They also found that Twitter can be exploited as a customer relationship management channel, as 19% of the entire population of tweets mentions an enterprise or a specific product. Recent researches by NM Incite and Edison Research have confirmed this trend, showing that 60% of social media users create reviews of products and services [6] and one in four users knowingly follow brands, products or services on social networks [20]. In this scenario, Twitter and other social networks stand as a great opportunity as well as a risky environment, where companies are exposed to brand hijacking (cfr. [21]) and other kinds of problems.

As a conclusion, an emerging research topic in this area is the so-called *crowdsourcing*, that is, the idea of involving a distributed group of people

in order to perform some tasks. In this sense, a significant example can be found in [22], that illustrates SocialEMIS, a tool that aims to enabling a worldwide collaboration for emergencies management, as well as in [23], that proposes CrowdSearcher, a novel search paradigm whose objective is the integration of generalized search systems with social platforms in order to capture users' opinions and preferences. Thereby, we can imagine a future in which social networks will play an increasingly important role.

2.2 A sociological perspective

At this point, we will adopt a different perspective, taken from sociology, in order to study the structure of online communities from a theoretical point of view and to derive real-world applications.

2.2.1 Social networks structure

Graphs, together with the sociometric and algebraic notation, are one of the three main ways to describe social network data mathematically. The graph theoretic notation scheme is "a model of a social system consisting of a set of actors and the ties between them" [24]. As usual, we refer to model as a simplified representation of a situation that captures and mirrors some, but not all, of the aspects of the situation it represents. When a graph is used to model social networks, nodes are used to represent actors, and edges the relations between the actors. There are in the literature numerous studies comparing the characteristics of social networks with those of other kinds of networks (e.g. the World Wide Web, peer-to-peer networks, food webs, etc.). For instance, Newman et al. [25] claimed that social networks differ from others for two reasons: the fact that the degrees of adjacent vertices in networks are positively correlated and their high transitivity. First of all, we must define the

concept of degree: the degree of a node is the number of lines that are incident with it [24]. In case of social networks, the degree simply represents the number of actors that are in relation with a given actor. From now on, we will use the concept of friendship as an example of relation between two people. Therefore, a specific individual has degree n if it has exactly n friends. It turns out that the degrees of two adjacent vertices (i.e. people) in the network are positively correlated if the two people are very likely to have quite the same number of friends. According to [25], this peculiarity, that does not belong to any other type of network, is due to the fact that social networks are usually divided into groups or communities. This is the same reason from which it descends clustering or transitivity: given a triple of actors i, j, and k, and the ties between them, the triad involving actors i, j, and k is transitive if whenever i is in relation with j and j is in relation with k, then also i and k will probably be connected [24]. It means that in a social network, the fact that i is a friend of j and j is a friend of k, makes it likely that i and k will be friends. With the term "likely", we mean that the observed clustering is greater than that expected for an equivalent random graph model. The same does not hold for other kinds of networks. The importance of such discussion resides in the fact that it applies also to the online users population. As an example, a 2009's research by Sysomos [26] has noticed the presence on Twitter of a lot of communities, that are collections of closely linked users (i.e. in relation with each other as *followers*), and thus we expect to find there the same previously defined features. Other studies showed the convergence between the social and technological networks, which constitute a huge sample of collective human behaviour and thus can be seen as a benchmark for proving sociological theories. For example, Watts and Strogatz pointed out that many social and technological networks have small path lengths and call them a "small-world" [27]. It means that in a small world most of the actors are just a few steps apart. The social psychologist Stanley Milgram has found the median path to be six steps long (and this is

where the expression "six degrees of separation" comes from) with a practical experiment conducted on a small scale [28]. Surprisingly, Leskovec and Horvitz, in a recent study, derived a social network from a quarter-billion user profiles on MSN Messenger and assessed that same number to be approximately 6.6, remarkably close to Milgram's result [29]. According to Watts and Strogatz, the reason behind the small world phenomenon is that the addition of a small number of random social links to a highly clustered network (like social networks, as discussed before) causes it to become a small world. A further research by Kwak et al. [17] focused the attention on the topological characteristics of Twitter and highlighted some small differences with respect to common human social networks. Firstly, the low level of reciprocity, i.e. mutuality in the relation that is mainly due to the fact that many celebrities or news media have a huge number *followers* (i.e. people that established a unilateral relation with them), which in turn are not followed by them. Secondly, the average path length between two nodes is 4.12. Once again, this can be explained by the fact that people set up relations not only for social networking, but also for seeking information. Moreover, Java et al. [12] computed a degree correlation of 0.62 that makes Twitter very close to the real social networks. Finally, Huberman et al. [30] proposed a slightly different concept of friendship on Twitter that solves the problem of low reciprocity. They defined friend "anyone who a user has directed a post to at least twice" and distinguished between two social networks: the one made up of *followers* and *followees* that is dense, and another one sparse and simpler but more influential in driving Twitter usage.

2.2.2 Social networks diffusion models

Back to the Milgram's experiment, it showed not just the presence of short paths, but also that people were able to find them. The first question that comes to everybody's mind is: how can that be? Some

theoretical models have been proposed but, of course, they could not have been proven until rise of social networking sites. A study by Liben-Nowell et al. [31] surprisingly demonstrated the adherence of such models to the friendship network of the blogging site LiveJournal with only slight adjustments. Anyway, information does not flow to just one target, but radiates in many directions at once. Therefore, we refer to this phenomenon as "social contagion". Some basic mathematical models claim that people's adoption of new or particular behaviours (e.g. the purchasing behaviour) is related to the behaviours of their neighbours in the social network [32]. Backstrom et al. [33] proved this thesis showing that the probability for a user to join an online group depends on the number of friends who belong to the group and on the way those friends are connected to one another.

The idea of social contagion expresses itself also in terms of spread of ideas and emotions. Chmiel et al. [34] collected over 4 million comments published on blogs, BBC discussion forums and the popular social news website Digg. With the aid of a machine-learning classifier, they categorized the whole set of posts according to their *sentiment*. They found many groups of consecutive messages (i.e. clusters) with the same emotional valence, longer than clusters from random distributions, thus demonstrating that the spread of emotional states actually occurs even in virtual environments. Likewise, a recent work by Kramer [35] has shown the existence of a correlation between the posts of some Facebook users and the subsequent posts of their friends. In particular, when a Facebook user publishes a message, the words he choose influence the words chosen by his friends up to three days later. Hence, he concluded that emotional contagion is possible also through communications that rely only on written language and indirect communications media.

2.2.3 Real-world applications

From the previous discussion, it follows that social networking sites mirror real social networks and thus can be exploited for several applications. As reported by Kleinberg [36], the design of numerous information systems were inspired by the idea of contagion, with the birth of the so-called gossip protocol, i.e. "a style of computer-to-computer communication protocol inspired by the form of gossip seen in social networks" [37]. Some researches in distributed computing have given rise to "epidemic algorithms", so called because information updates are spread between hosts in a manner similar to the way that a viral infection spreads in a biological population. Furthermore, other possible applications concern the use of microblogging for *electronic word of mouth* (*eWOM*) branding. In fact, as said before, Jansen et al. claimed that customer brand perceptions and purchasing decisions are influenced by online communications and social networking services. Moreover, the six degrees of separation (or even less in case of Twitter) confirm the fact that tweets published by people with a limited amount of *followers* can reach a huge audience, as Kwak et al. stated. They also showed that after the first *retweet*, a second, third, up to fourth one occurs almost instantly, and all this follows directly from what we have explained above. It turns out that a sociological background constitutes the basis over which new theories and applications in this field of research can be developed.

2.3 Automated comprehension of text

2.3.1 Background on Natural Language Processing

Natural Language Processing (*NLP*) is "a theoretically motivated range of computational techniques for analysing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications" [38]. According to Liddy [38], recent psycholinguistic researches suggest that human language processing is a dynamic analysis conducted at different levels simultaneously (synchronic model of language) and not in a sequential manner (sequential model). However, the key point is that all the levels contribute to meaning. We present here the classification of these different levels (all the definitions and explanations are taken from [38]).

Phonology

Phonology is the level focused on the analysis of speech sounds within and throughout words, which is typical of *NLP* systems that accept spoken input. Such systems analyse and encode the sound waves into a digital signal that is then decoded by means of various rules or by comparing it to the particular language model in use.

Morphology

Morphology is the level regarding the structural nature of words, in particular it deals with morphemes, which are the smallest linguistic units with a given meaning. For instance, morphology tells us that the

word “unlikeable” is made up by 3 different morphemes whose meaning remains the same across words: the prefix “un” signifying “not”, the root and free morpheme “like” signifying “affection for something or someone” and the suffix and free morpheme “able” signifying “can be done”. Another example is the suffix “-ed” added to a verb, which means that the action took place in the past. An *NLP* system is designed to identify each morpheme and understand their meaning in a similar way as humans do.

Lexical

The lexical stage is the one at which the meaning of each single word is examined. For instance, the assignment of each word to its part-of-speech tag pertains to the word-level understanding. This level might require a lexicon, which could be either simple, consisting only in the words and the part(s)-of-speech, or it may also include information such as the semantic class of the word, the arguments that it takes or its sense(s) definitions in the given semantic representation.

Syntactic

The syntactic stage is the phase in which the focus is on the grammatical structure of the sentences, revealed by the analysis of the words comprising them, thus making essential the presence of both a grammar and a parser. A representation of the sentence in which the structural dependencies among words are highlighted is the output of the syntactic processing. The syntax level is the one that mostly contributes to understand the meaning of a sentence, because the order of words is what conveys most of it. For instance, consider the two syntactically different sentences “The book is on the table” and “The table is on the book”: they have the same exact words but in different order, thus resulting in opposite meanings.

Semantic

Semantic processing looks at the interactions among word-level meanings within each sentence so to identify its possible meanings. For polysemous words, this level can also envisage the selection of one of their meanings thus performing what is called the semantic disambiguation. For example, "Milan" can be interpreted as the city or can be the noun of the popular Czech writer Milan Kundera. If additional information are needed, the semantic instead of the lexical level would perform the disambiguation.

Discourse

At discourse level, the analysis is performed at the granularity of group of sentences. This does not trivially mean that the global *sentiment* is evaluated by aggregating the results of each single sentence, processed singularly. On the contrary, each document is considered as a set of sentences mutually connected. Hence, the polarity can be estimated only by looking at the content as a whole. This may involve the application of various techniques:

- *Anaphora resolution:* it is the substitution of particular terms or expressions (e.g. pronouns) with their referential element.
  E.g.: "Alice plays the piano. She doesn't know how to play the violin." "She" will be replaced with "Alice".
- *Discourse/text structure recognition:* it refers to the identification of the roles of different parts in a text.
  E.g.: an essay is usually made up of three different parts, i.e. the introduction, the body and the conclusion.

Pragmatic

We act at pragmatic level whenever we make an implicit use of an ensemble of contextual information that is necessary for capturing the meaning. In this case the main problem for an *NLP* tool is that the context is something of which a human reader is aware even if it is not explicitly mentioned or encoded in a text and thus cannot be extracted from it. For instance, when we encounter the word "man", we exploit contextual information in order to interpret if it refers to the whole human species, to the males of the human species (as opposed to woman) or to the adult males of the human species (as opposed to boy).

The key point is that each level of analysis is important as it contributes to the formation of what we call "meaning"; thus, the more levels of language an *NLP* system utilizes, the more capable it is. However, current *NLP* systems tend to focus on just few steps at the lower level of language processing. This is due to the fact that some applications do not require interpretation at the higher levels and to simplicity reasons.

2.3.2 Sentiment analysis

Sentiment analysis (aka opinion mining) refers to "the application of *natural language processing*, computational linguistics, and text analytics to identify and extract subjective information in source materials. Generally speaking, *sentiment analysis* aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document." [39]. This subjective information can appear in two different forms: it can be evident, as it is explicitly named (e.g. "What a wonderful car!") or it may be left implied (e.g. "I've just bought a new television but it doesn't work") [40]. Moreover, an explicit sentence can report an absolute opinion (e.g. "The image quality of this

camera is good") or a comparison (e.g. "The image quality of camera X is much better then camera Y") [41].

There are many orthogonal classifications of *sentiment analysis* techniques. For instance, they can act at different levels [41]:

- *Document level:* the goal is to analyse a text and to try to identify the overall tone expressed. The final result is a classification according to the polarity, in terms of positive or negative orientation.
- *Sentence level:* it can be split into two subtasks. First, it has to be identified whether the sentence conveys an opinion (i.e. it is subjective) or a mere description of a fact (i.e. it is objective). Subjective sentences, as opposed to the objective ones, contain something that cannot be proven or refuted (e.g. emotions, opinions and speculations). However, it must be pointed out that a subjective sentence can be true and, likewise, objective does not imply true [40]. The second task is to classify a subjective sentence and determine the attitude of a speaker.
- *Feature level:* it is the most fine-grained analysis and the hardest to perform. It aims to state not only the global polarity of a document, but also what the author thinks about all the specific characteristics of an object or topic mentioned in the text. This objective can be split into four subtasks: the identification of relevant entities and of the features discussed by the author, the evaluation of the orientation of opinions and the production of a summary.

In another categorization, *sentiment analysis* tools are divided according to the specific approach applied in order to identify the semantic orientation of words [40]:

- *Lexical:* it uses a lexicon, manually or automatically constructed, and applies linguistically derived heuristics. Moreover, as *sentiment analysis* is an application of *NLP*, different techniques can perform analyses at the previously discussed levels of the language.
- *Statistical:* based on statistical techniques, such as gathering knowledge from the co-occurrence of a new term with one whose polarity is already known.
- *Machine learning:* training documents play a fundamental role as a source to extract useful information about the polarity of terms. Several specific techniques can be exploited (e.g. latent semantic analysis, support vector machines, etc.). A further classification can be done between the supervised and unsupervised approach [42]:
    - *Supervised (document classification):* a huge training set of human-coded documents is employed to derive the likelihood of some pre-defined category labels that have to be assigned to new documents.
    - *Unsupervised (document clustering):* there is no need for human intervention or labelled documents at any point in the whole process. The aim of text clustering is to group text documents such that intra-group similarities are high and inter-group similarities are low.

Finally, we can distinguish different ways of modelling human mood:

- *Monodimensional:* there are two variants [43]:
    - *Binary:* a subjective sentence can be labelled either as positive or negative.
    - *Multi-scale:* the overall *sentiment* is estimated in terms of a Likert scale.

- *Multidimensional:* it measures mood in terms of several dimensions according to the "Profile of Mood States" (POMS), a psychological rating scale. It consists of 3 versions: the POMS Standard, POMS Brief & POMS-Bipolar [44]. Examples of affective states measured by POMS are tension-anxiety, vigour-activity, depression-dejection, etc.

As a conclusion, it must be pointed out that each technique has its pros and cons and that it is quite impossible to create an effective general purpose *sentiment analysis* tool. As reported by [41], "generic engines and algorithms perform much worse than applications meant only to analyse particular types of text", because opinions are expressed with an informal language and the way in which a sentence is built, that is, its structure is strictly related to the specific community. The extraction of *sentiment* from a text is a strikingly tricky task: as a matter of fact, even humans are often in disagreement. Furthermore, the shorter the string of text, the harder it becomes.

2.3.3 Applications

According to Pang and Lee [45] and Westerski [41], *sentiment analysis* has many application. We can cite for example:

- *Advertisement placement:* the current trend in commercial campaigns is toward targeted advertising, that is, the attempt to reach potential customers rather than a generic and vast audience. For instance, a topic-based mining technique can decide to display the advertisement of a specialized computer equipment next to its review on a tech forum. But what if the article releases a negative review of the product? The advertisement should obviously not be displayed and an opinion mining tool may guess this fact basing on the polarity of the article.

- *Business and government intelligence:* manufacturers can be interested in people's personal views and judgments of their products characteristics. Opinion mining tools can allow companies to get ahead of their competitors and swiftly react to customer needs.
- *Opinion search and retrieval:* same as above, but from the customer's perspective. Customers might use opinion search engines in order to look for opinions about products.
- *Understanding the dynamics e-communities: sentiment analysis* can help to explain the trend of certain e-communities (e.g. the success of Facebook over MySpace) by identifying how negative emotions affect the evolution of social networks discussions [39].

2.3.4 Sentiment analysis tools

Finally, we provide here a list of well-known *sentiment analysis* tools:

- *LIWC (Linguistic Inquiry and Word Count) [46]:* it has two central features: the processing component and the dictionaries. The processing component analyses the text by searching each word in the dictionary file and by extracting all the possible part-of-speech tags associated to the specific word and its emotion category (e.g. the words "maybe", "perhaps", or "guess" are representatives of the category "tentativeness"). After going through all the words in the text, LIWC would calculate the percentage of each LIWC category. The LIWC output, then, lists all LIWC categories and their rates of usage.
- *OpinionFinder [47]:* its main functionality is the identification of the words that convey a positive or negative *sentiment* in a given sentence. It runs in two modes, batch and interactive. In batch mode, OpinionFinder takes a collection of documents to analyse.

Moreover, users can query online news sources and process documents also through the interactive mode front-end. It is composed of two parts: the first performs general purpose document processing (e.g. tokenization and part-of-speech tagging). The second part performs the subjectivity analysis through a Naive Bayes classifier that exploits a variety of lexical and contextual features. The results are returned to the user in the form of SGML/XML mark-up of the original documents.

- *Summize4 [18][48]:* it was a popular service (purchased by Twitter in 2008) for searching tweets and assigning them an overall *sentiment* rating for a given period using a five-point Likert scale. The five classes of *sentiment* are wretched (i.e. purely negative), bad (i.e. mainly negative), so-so (i.e. mediocre or balanced *sentiment*), swell (i.e. mainly positive) and great (i.e. purely positive). Summize uses a lexicon of 200000 uni-grams and bi-grams (i.e. sets of two terms) of words and phrases that have a probability distribution to determine the *sentiment*. Summize employs a multinomial Bayes model as classifier, which selects the best class (i.e. the one with the highest probability) in a winner-takes-all scenario.

2.4 Prediction from social media

2.4.1 Motivation

In recent years we have assisted to the proliferation of social media monitoring and analytics tools (e.g. BuzzMetrics by Nielsen Online, Alterian's SM2, Radian6, etc.). Of course, their usefulness resides in the assumption that social media platforms truly reflect public opinion. The same holds for the idea of predicting future trends, a topic that has captured the interest of many scholars (cfr. [49]). Therefore, we must

answer to this question: are online *sentiment* and public opinion strictly correlated? A possible answer was suggested by Brendan O'Connor et al. in [50]. What they did was to take classical polls and surveys as an indicator of the public opinion about two main topics: consumer confidence, i.e. how optimistic the public feels about the health of the economy and their personal finances, and politics. In the meanwhile, they monitored tweets that contained some accurately selected words which were significant for their topics of interest. All these messages were processed by a simple lexicon-based *sentiment* classifier, OpinionFinder, in order to assess if they expressed a positive, negative or neutral opinion. The measure adopted to estimate the *sentiment* on a particular topic in a certain day was the ratio of positive versus negative messages. They found that the two time series of public opinion and online mood were reasonably correlated. Therefore, if a relatively simple *sentiment* detector based on Twitter data can replicate consumer confidence and presidential job approval polls, it turns out that the *sentiment* extracted from social media sites is a good indicator of public opinion or at least as good as the widely-spread traditional surveys.

2.4.2 The early Web 2.0 days

The idea of the existence of a latent correlation between *User Generated Content* and real-world outcomes is quite old and dates back to early 2000's, when social media were not so spread. For example, Gruhl et al. [51] addressed the problem of discovering a possible connection between online chatter and the purchasing behaviour of customers. In particular, they focused on the field of books sales. At that time, neither Facebook nor Twitter did exist, thus they had to rely on a slightly different data set: the IBM WebFountain project (a huge collection of online unstructured material) for online postings data, and Amazon for sales data. They analysed the sales rank time series of all the books that became part of the top 200 most sold books at least once during the

observation period, in order to identify a subset of books which had a spike in their sales. After that, they created a user interface that allowed to easily submit a hand-crafted query, with the purpose of retrieving blog mentions about a specific book. Finally, they plotted the blog mentions and sales rank trend and analytically computed the cross-correlation of the two time series. The results of this work lead to the conclusion that if there is a spike in the sales rank and there are lots of blog mentions about the book, then the blog mention tends to have a spike that is correlated and often anticipates the other spike. This is probably due to the fact that many of the reasons that cause a spike in the sales also have the potential to generate a spike in blog mentions, despite of some inevitable biases (e.g. discounts, promotions, privileged positions on the website, etc.). Starting from this point, they built some complex models to predict the behaviour of sales ranks, but none of them seemed to be very effective. However, they succeeded in building a good performing predictor for future spikes in sales ranks, which takes the blog data as unique input. Of course one of the major drawbacks in this work was the lack of a voluminous online chatter. In fact, as said before, the information sources considered were limited to blogs, the only *Web 2.0* examples available at that time.

Another example of forecasting model which does not base itself on social media data has been defined by Hyunyoung Choi and Hal Varian, Google's chief economist, in [52]. They used Google Trends, a service that shows how often a particular search-term is entered relative to the total search-volume across various regions of the world and in various languages, in combinations with other specific data sources, to feed a simple linear regression model. This model was used to estimate future time series values (with a time step of one month) based on past observations and Google Trends data related to the first week of the considered month. They applied this methodology to many different forecasting scenarios, such as automotive and home sales, travel, etc. The

results clearly showed that Google Trends variables are statistically significant and that simple models that include these relevant variables always outperform models that exclude them (with an improvement in terms of *Mean Absolute Error* often greater than 10%). Hence, it can be concluded that the volume of queries on a given topic provides useful information for predicting future trends.

Finally, Mishne and Glance [53] moved one step forward by introducing the concept of *sentiment* in their analysis. They started from Gruhl's considerations and applied the same methodology in the domain of movies. They built many sets of variables, one with the simple chatter volume and the others with *sentiment* indicators (e.g. count of positive posts, ratio of positive versus negative feedbacks, etc.). In line with what Gruhl et al. had already proved, their results showed that there was good correlation between references to movies in weblog posts and the movies financial success. Furthermore, as the number of positive references correlates better in the pre-release period, they demonstrated that *sentiment* can be exploited for predictive purposes.

2.4.3 From blogs to social media platforms

The birth and huge diffusion of social networks represented a turning point in this research area, as the collection of huge data sets of *User Generated Content* became possible.

Asur and Huberman [54] demonstrated how chatter from Twitter could be easily exploited to forecast box-office revenues for movies. They split their analysis into two parts; firstly they focused on a simple indicator, i.e. tweeting rate, then they showed how the introduction of a *sentiment* variable could provide further improvements in the forecasting power. They started from a basic assumption: movies that are well talked about, will be well-watched. In order to assess the validity of such hypothesis they crawled Twitter data related to some movies for a three weeks

period (from the week before the movie was released to two weeks after). As a first attempt, they examined the correlation between the percentage of URLs and *retweets* of promotional materials with the box-office performance, but unsuccessfully. Better results were obtained with a simple linear regression model, which surprisingly outperforms the Hollywood Stock Exchange Index (HSX) that has been shown to be a good predictor of real box office results. The set of independent variables includes only two basic components: the tweet-rate time series and the count of theatres a particular movie was released in. As we said before, this model outperforms the HSX-based model according to the adjusted R^2 and the *p-value* indicators. Finally, they tried to investigate the importance of the so-called *sentiment* in predicting the revenues after the release. They classified tweets in terms of positive, negative or neutral *sentiment*, with an accuracy of 98%. After that, they defined a measure, called PNratio, which is simply the ratio between the number of positive and negative tweets. The idea behind this is that positive tweets can be considered as recommendations by people who have seen the movie and are likely to influence others from watching the same movie. Once again, they built another model by performing a linear regression on the second weekend revenue, with tweet-rate and PNratio as independent variables. The result was an improvement of 0,1 in the adjusted R^2 from the model based on tweet-rate time series to the one using the PNratio as an additional variable.

Another well-known topic in this research area is related to political forecasting. This interest has been originated by the successful campaign of Barack Obama in the U.S. presidential elections, which designated social media websites as the leading platforms to engage voters. In fact, Obama's social-networking website mybarackobama.com, also known as MyBO, helped him in gathering information about potential voters' political leanings and in getting volunteers involved in the campaign (cfr. [55]). Many researchers demonstrated how very simple indicators, such as

the number of Facebook supporters of a candidate or even the count of mentions in the press, could sometimes outperform traditional polls. Tumasjan et al. [56] extended these studies by taking advantage of the birth of new social network platforms, such as Twitter. In particular, they concentrated their analysis on the German political landscape and tried both to evaluate whether tweets reflect the offline political *sentiment* in a meaningful way and to figure out a way to forecast the outcome of elections. They gathered and examined all the tweets mentioning one of the six major parties or their politicians that were published during the one-month period prior to the elections. With respect to the first goal, they analysed this data set by using the text analysis software LIWC2007, which calculates the degree to which a text sample belongs to a specific psychological category (e.g. future orientation, anxiety, tentativeness, etc.). Such analysis revealed the existence of a great correspondence between the psychological profile of a politician delineated on Twitter and his behaviour during the campaign. Moreover, they observed that joint mentions of political parties accurately reflect prevailing or upcoming political ties. Therefore, they concluded that tweets convey many nuances of the election campaign. Finally, they demonstrated that a very basic indicator, such as the tweet volume per party, has a predicting power that is comparable with the one of traditional election polls. They also pointed out some serious shortcomings that can be seen as general limitations for all the researches of this kind, such as the fact that usually a small number of users generates the great majority of tweets and the implicit difficulty in the process of automating the categorization of very short messages as tweets are. Similar results were obtained in the Senate race between Sharron Angle and majority leader Harry Reid by using Crimson Hexagon, a Harvard-developed analytics tool (cfr. [57]).

2.4.4 Exploiting the sentiment

Until now we have seen examples of prediction models that rely on social media content as it is, i.e. without any pre-processing. These kind of model can perform well in some cases, but there are lots of situations in which an input refinement is required. *Sentiment analysis*, as defined in the previous chapter, should play a fundamental role in the construction of a reliable predictive system and may indeed improve forecasts, as suggested by [53] in the case of weblogs.

In order to better clarify this point, let's turn back to the previous example of political elections. Both Chung and Mustafaraj in [58] and Metaxas et al. in [59] expressed their doubts and concerns for Tumasjan's work and showed how their great results were not repeatable in a different environment. Specifically, they both focused on the 2010 U.S. Senate special election in Massachusetts and applied the same methodology, as it was proposed by Tumasjan, but unsuccessfully. They suggested that the share of mentions a candidate receives is not enough to predict an election outcome and that such discrepancy is probably due to the fact that also the *sentiment* of a message must be taken into account, as many tweets express unfriendliness rather than support for a candidate. Furthermore, they both underlined the inadequacy of some well-known *sentiment analysis* tools (e.g. OpinionFinder and SentiWordNet), which are rarely able to reach good levels of accuracy (just a little bit more of a random classifier in many cases). The problem is that these tools are based on lexical or syntactic techniques while we are moving towards semantic analysis of *sentiment*. Finally, Metaxas et al. claimed that there are significant differences in the demographics between likely voters and users of social networks (as it is evident from a Pew research [60]), and thus the latter ones cannot be considered a representative and unbiased sample of the population.

Anyway, in spite of all the problems stated before, there are examples that show how the application of effective *sentiment analysis* techniques can improve the predictive power.

In fact, these ideas have been also applied to the prediction of recent social phenomena. According to [61], it has been demonstrated that feeding a supercomputer with news stories could help to predict major world events. A fundamental part of the process is the assignment of a mood (e.g. good or bad) and a location to which the news refers. The results are astonishing: as an example, they plotted the media *sentiment* around Egypt in early 2011, just before the resignation of President Mubarak. What they found is that the tone of the media fell dramatically to a low only seen twice before in the preceding 30 years, and this fact could be interpreted as a clear signal that a turning point was approaching.

Finally, let's come back once again to the problem of political forecasting. Choy et al. [62] took a quite revolutionary approach in order to overcome the problem of having a limited number of social media users, that may not be so representative of the entire population. They started from an analysis of the Singapore political environment, which is theoretically a two-party system but in practice has been dominated by People's Action Party for half a century and where online debating has begun to play an important role, as discussed in [63]. However, a negligible share of the Singaporean voters used social media at the time of the study. Therefore, they suggested as a smart solution what they called "census correction". First, as usual, they extracted tweets published during the campaign period. Second, they performed a *sentiment analysis* on the data set with a customized corpus, i.e. an ad hoc tool specifically developed to avoid the typical problems of an automated general purpose system (e.g. the localized version of English). Third, in order to better estimate the votes and bypass possible biases in the data, they collected statistics about the distribution of age groups in

the population, computer literacy and percentage of support for a specific party or candidate in a certain age group. Finally, they combined all these pieces of information with the previously extracted *sentiment* and they computed, with the aid of simple formulas, the share of consensus that each candidate would have got at the presidential elections. The *sentiment* value, that is, the aggregated positive and negative emotions from the tweets, made evident that Tumasjan's approach would not suffice to justify the actual outcomes. In fact, all the candidates got a comparable level of support on social media. The suggested methodology, instead, succeeded in predicting the correct order of preference, except for the first two candidates, whose actual shares differed only for a small margin of votes: 35,19% versus 34,85%. Nevertheless, also in this borderline case, the estimated outcome realistically modelled the real one, with a predicted difference in share of just 1,1%. Moreover the authors suggested that such forecast error may be due to the fact that their equations do not model the concept of swing voters, i.e. people who has no allegiance to any political party and whose unpredictable decisions can swing the outcome of an election one way or the other.

As a conclusion, we can say that the mere number of successful examples justifies the great interest in this research area. Anyway, in any research work a special attention must be posed to the data processing phase, because some currently employed *sentiment analysis* techniques seem to be not very effective and may lead to misleading results.

2.5 An introduction on algorithmic trading

Within the financial market, the scenario over which we are focusing our attention is the one of trading equities, that is the buying and selling of

companies stock shares through one of the stock exchanges such as the New York Stock Exchange (NYSE) [64].

A natural evolution of equity trading is the so-called algorithmic trading (also known as quantitative trading or black-box trading) which is its automated version, where algorithms are designed to perform the trading activities by choosing the right time to place the orders along with the number and the price of the stocks to buy or sell [65].

A further yet extreme evolution of algorithmic trading is High Frequency Trading (HFT) which is distinguished by a much shorter term view, that is reflected in a larger number of intricate buying and selling orders on the selected financial instruments which can be either equities, futures, ETFs, derivatives or currencies. More in deep HFT strategies can be quantitative investing strategies deployed at high speed or some specifically tailored strategies to work with market microstructure. What HFT generally exploits are small intraday variations in prices, which do not influence long-term investors [66].

Quantitative trading can be seen as the implementation of a trading strategy in such a way that it could be thought as an algorithm, thus enabling its automatic execution.

Financial as well as non-financial experts with a scientific background conduct thorough researches to find the set of securities they will trade and what data they will need to implement the trading strategy. The people behind quant trading strategies are the so-called quants or quant traders and as it is often the case they have received scientific yet non-financial education and they may be experts in physics, statistics, mathematics, and computer science [67][68][69].

In order to understand the practical significance of quantitative trading in the stock market arena is important to notice that, as indicated by Aite Group in early 2009, more than 60 per cent of all U.S. equity transactions were due to short term quant traders [70], while Rob Iati, a Partner of TABB Group, writes that despite high-frequency trading firms represent just about 2% of the twenty thousands trading firms

operating in the U.S. markets they account for 73% of all U.S. equity trading volume [71].
While considering the European market, transactions generated by computers based on mathematical formula accounted for 45 per cent of volume in Boerse's Xetra electronic order-matching system in the first quarter of 2008 and just a year earlier the same percentage was 33 [72].

What distinguishes a quantitative trading strategy from a discretionary trading strategy is the total absence of arbitrariness, that means removing decisions driven by emotions, indiscipline, passion, greed and fear encouraging an analytical and systematic approach [67].
Going a step further we can distinguish between two broad classes of strategies: the strategies that attempt to generate returns by expertly sizing and timing of diverse portfolio holdings and the strategies that slightly improve on the performance of an index. The formers are called alpha strategies and are the ones tracked by quants and the latter are called beta strategies that are in the focus of discretionary trading (which deals also with quantitative analysis). Therefore, quants pursue alpha or returns that are independent of the direction of any market in the long run [67].
Algorithmic trading can be employed in any investment strategy, such as trend following (or momentum), delta neutral strategies, mean reversion (or counter-trend), arbitrage, and a vast set of other proprietary strategies [65][67].
Trend following is by far the simplest technique to understand as it tries to exploits the moves that sometimes occur in the price of a given financial product. The strategy aims to take advantage of a market trend, going long (that is buying) or short (that is selling) in a market, with the objective of gaining profits from the ups and downs of price of the given financial instrument. It is worth noting that this has always been and it is likely to remain one of the most important ways in which traders go about their business [67].

Delta neutral strategies are devoted to keep unchanged the value of a portfolio of related financial securities taking advantage of small changes in the value of the underlying securities [73].
Mean reversion strategies are concerned in the buying or selling of a certain security whose recent performance has greatly differed from the historical averages [74], that is, when the current market price is lower than the average price, the stock is considered attractive for purchase with the aim of selling it back at a higher price and conversely when the current market price is above the average price, the price is expected to fall. In general, deviations from the average price are expected to revert to the average [75].
Strategies based upon arbitrage can be applied either when the same asset does not trade at the same price on all markets or when two assets with identical cash flows do not trade at the same price. These strategies take advantage of the price difference by simultaneously buying a financial product in one market and selling it in another market [76]. Statistical arbitrage is a subset of these strategies and a classic example of it is pairs trade: consider two stocks, one for company A and the other for company B, with similar market capitalizations belonging to the same industry, with similar business models and financial status. Company A is included in a major market index, an index that many large index funds are tracking, while company B is not included in any major index. It is likely that the stock of company A will outperform shares of company B thus a pairs trade will be to simultaneously open a short position on company A's stock and a long position on company B's stock [67].
It is worth noting that there are various techniques, such as VWAP or TWAP, designed to get an order done without creating any detectable market impact. The volume weighted average price (VWAP) strategy aims at placing an order by monitoring the trading volume on a given time frame with the goal of reducing the trade impact on the market, thus preventing any appearance of abnormal trading activity, which

could damage the price at which the order is executed as other traders realize what is happening. The time weighted average price (TWAP) technique trades based on the clock, allowing the partitioning of a trade over time, and it is applied in small illiquid stocks [77].
Finally, there is a broad spectrum of black-box algorithms and techniques that are proprietary of various financial firms. For instance, Credit Suisse's Advanced Execution Services unit, which serves major hedge funds and other buy-side clients since its creation in 2001, developed a vast set of proprietary strategies [78]. One of these, named "Guerrilla", is an agile trading strategy capable of sourcing liquidity amid challenging market conditions. "Using real time market activity to intelligently adjust trading behaviour, Guerrilla 2012 takes full advantage of extreme, unexpected, price moves that work in the client's favour while constantly guard against movements that work against them" [79].
More than 20 different financial firms have developed and advertised their own liquidity seeker algorithms with stealthy capabilities that are available for their clients [80].
According to Rishi Narang, a quant systematically applies an alpha seeking investment strategy that was specified based on exhaustive research. The peculiarity of a quant resides in how an investment strategy is conceived and implemented. Quants and discretionary trades rarely differ in what their strategies are actually doing (for instance, consider pairs trading and statistical arbitrage that both kind of traders can implement) [67].

The base components of a live, production algorithmic trading system are the ones that decide which securities to buy and sell, how much, and when. An alpha model is the first component and it is devoted to predict the price of the selected financial instrument with the goal of achieving the largest possible profits. Correspondingly, a risk model has to be put in place and designed for limiting the odds that may generate losses. As almost any trading transaction costs money, independently from the

expected profit or loss of the trade, there is also the need to limit the expenses related to the execution of the trades that are performed to change the current portfolio structure; this is exactly what the transaction costs model does. These three models precede the portfolio construction model, which is designed to optimize the portfolio structure in such a way that the profits are maximized while the risk and the transaction costs are being kept as low as possible [67].

Finally, the portfolio construction model interacts with the execution model that is in charge of executing the trades for building the new portfolio while considering important factors such as the market liquidity and the relevance of each transaction.

Within a quant trading system, the alpha model is where much of the research process is focused, as it is the part devoted to search for the profits. Alpha is generally used as a way to quantify the skill of an investor in terms of the portion of the investor's return not due to the market benchmark, that is the value added or lost by the investor. Alpha models are therefore the quant's added skill to the investment process in order to gain profits and may be named forecasts, factors, alphas, strategies, estimators, predictors or exposures because they are designed to anticipate the future with enough accuracy. "The alpha model is the optimist, focused on making money by predicting the future" writes R. Narang in [67].

The small number of trading strategies that we have briefly explained can be implemented in many ways, though we can distinguish between theory driven ones from data driven ones. The former are based upon theories built from observations of the markets that are subsequently tested against real market data. For example, "cheap stocks outperform expensive stocks" is a theory that explains the existence of countless value funds. The latter are focused on detecting patterns in the data by exploiting data mining and statistical techniques.

A further classification of theory driven strategies is made between the ones that exploit price-related data (for instance prices of various

instruments or other information that comes from an exchange such as trading volume) from the ones that make use of fundamental data. Trend following and mean reversion strategies are based on price-related data while value (or yield), growth, and quality based strategies are constructed on fundamental data such as price to earnings (P/E) ratios, price/earnings-to-growth (PEG) ratios, analysts' earnings estimate revisions, debt-to-equity ratios [67].

As a conclusion, it is worth noting that, in the last 5 years, firms such as Thomson Reuters, Dow Jones and Bloomberg began to format financial market news to be read and traded on via algorithms offering unprecedented possibilities. *Natural Language Processing* (*NLP*) and *sentiment analysis* techniques are now being used extensively to process financial news coming from various sources (also including social networks in the more recent trend) thus creating a feed into some specific autonomous systems which then trade on the news [81][82].

2.6 Using the sentiment to predict the stock market

The belief that market prices can be predicted, at least partially, started when the first critics to the *Efficient Market Hypothesis* (*EMH*) were beginning to take place [83][84].
The *EMH* is associated with the idea of a random walk, which is typical of a price series where each price is a random result from the preceding prices. From another perspective this is the same of saying that the price change of a given day reflects only the news of that specific day thus being independent from the preceding price changes, assuming that the news flow is not interrupted and news have a direct impact on stock prices. Because of the unpredictability of the news, price changes must be

unpredictable and random. This can be summarized as "prices fully reflect all known information" [84].

In [83], Nofsinger deals on how social dynamics influence the financial world. He states that consumers as well as investors and managers are somehow driven by the social mood, which is affecting their decisions. Furthermore, he supports the idea that "The cues we obtain from others influence our own opinions. A shared attitude, or social mood, is thus propagated."

"Forgas [1995] presents a model that measures to what extent people are likely to rely on emotions in decision-making. He argues that decision characteristics like risk and uncertainty are factors in whether feelings play a role. The greater the complexity and uncertainty of a situation, the more emotions influence the decision."[83]

The idea of *sentiment* exploitation to gather relevant information regarding a given stock is state of the art in equity investing [85] and in Foreign Exchange Market (known as FOREX) [86].

There are news analytics products that convert news, for example from Dow Jones, into *sentiment* scores that can be utilized by quantitative analysts to make investment decisions. Leaders in this sector are Ravenpack, Thomson Reuters and Alexandria.
This approach to quantitative trading lead to two main technical difficulties: in the first place, the extraction of the *sentiment* from a given text fragment raise the problem to efficiently adopt *Natural Processing Language* (*NLP*) techniques, secondly, once the *sentiment* engine is in place and working with reasonable effectiveness, the mood change has to be linked with a given stock price change with a properly defined model.
Prior to the massive widespread of social networks, that started around 2008, early works with the objective of finding some predictive power as well as some useful correlation between whatever source of *sentiment* and

the financial markets, were mainly targeting stock message boards and news websites [87][88]. The market's reaction to a given news story is deeply investigated in [89].

As behavioural economics provided proofs that financial decisions are significantly driven by emotion and mood [83], in the last 4 years, some authors began to study if the public mood is correlated or even predictive of economic indicators by taking online social networks as a source of "universal mood" [90][91][92][93].

In the last decade, different machine learning techniques, such as Genetic Algorithm, Neural Networks, support vector machine (SVM), support vector regression (SVR) have been adopted to predict prices or the direction changes in the stock market or Foreign Exchange currency market.

A method for positively or negatively labelling news stories about publicly traded companies according to their apparent impact on the performance of the company's stock is examined in [89], where it is shown the existence of many lexical markers for bad news but none for good news. A news is defined as positive if the corresponding stock price rose for 10% at least and as negative if the stock declined for 7.8% or more. News satisfying a more constraining condition are then considered so to better link those news to the corresponding stock price changes. All words that appeared at least 60 times in the corpus are used as the features set and then each text is represented as a binary vector reflecting weather or not a feature is present. The categorization methodology consist in the selection of the 100 features with highest information gain in the training corpus and then using a linear support vector machine (SVM) to learn a model.
When the approach is to match each story with the price change of the corresponding stocks from the market closing price the day before the publication of the story to the market opening price the day after the

story, training on the entire news corpus ranging from 2000 to 2002, and testing on the 2003 corpus a linear SVM and various other learners such as Naive Bayes and decision trees, yielded an accuracy above 65%.
It is shown that a number of features are clear markers of negative documents, for instance documents containing terms such as "shortfalls", "negative" or "investigation" are almost always proven to be negative. The 77% of the examined news that have been classified as negative are really negative.
In [89] it is also explored the matching between news and the stock price change from the day after the publication of the news to the one two days after the publication of the news. The accuracy in this case is just above 52% thus bearing out the so-called *Efficient Market Hypothesis*: "prices fully reflect all available information".

Another paper that leans towards the *Efficient Market Hypothesis* is [87] where the relationships between Internet message board activity and abnormal stock *returns* and trading volumes are examined. This study focuses specifically on the Ragingbull.com discussion forum, an extremely popular site during 2001 and still in use today, where objective measures of investors' opinions can be retrieved. Despite the premise, one of the interesting findings of this work is that, on days with abnormally high message activity, swings of investor opinion are related with abnormal industry-adjusted *returns* and with abnormally high trading volume for stocks in the Internet service sector.
An event study was performed to determine the impact of high-message-volume days on securities' prices and trading volume by looking at the industry adjusted *returns* and abnormal volume around days with abnormally high number of posts. *Returns* following abnormal Internet message-board activity are statistically insignificant although days characterized by relevant positive opinions and atypical message board activity follow positive *returns*, and unusual message-board activity is coincident with abnormal stock *returns*.

Furthermore, a Vector Autoregression (VAR) analysis revealed that a linear one-day lagged time-series model paired with message-board data is not useful in predicting the *returns* of Internet service sector stocks, thus aligning this results with market efficiency, though the authors admit that with the proposed methodology it is impossible to determine whether daily activity on the message board causes or is the result of abnormal *returns* on the stock.

It is worth noting that as year 2000 coincided with the climax of the dot-com bubble that culminated with the decline of the XLK ETF (that is a technology sector ETF that included many of the Internet companies of that time) at the end of the year, this could had a non-negligible effect on the research done in that period.

Another academic work that studied the unpredictability of stock *returns* is [88], although it confirmed with significant evidence that messages posted on Yahoo! Finance and Ragingbull.com help to predict market volatility.
Naive Bayes is used as the "bag-of-words" approach for classifying messages in order to obtain buy, hold or sell signals for each one of them. Then, exploiting correlations and VAR models, it is shown that message boards can be helpful in predicting trading volume as well as the market volatility.
Furthermore, it states that stock message boards contain financial relevant information even though it is much noisier and less reliable than that conveyed by classical financial newspapers, such as the Wall Street Journal.

Alexandre d'Aspremont and Ronny Luss, when working at the Operations Research & Financial Engineering Department, examined in [94] the use of support vector machines to classify news articles with the purpose of predicting intraday price changes of financial instruments, thus exceeding the limitations of time series models such as

Autoregressive Conditional Heteroskedasticity (ARCH) and Generalized ARCH (GARCH) that have been developed to forecast volatility using asset *returns* without considering a key source of market volatility, i.e. financial news [88].
By means of support vector machines and text classification techniques, given the text of a press release (such as Reuters and The Wall Street Journal) and past absolute *returns* (i.e. volatility), it is predicted whether or not an abnormal *return* will occur in the next dozen of minutes; however, the direction of *returns* is not found to be predictable. The performance was greatly improved by the adoption of multiple kernel learning technique.

The *sentiment* gathered from Yahoo! Finance message board is modelled to be conditionally dependent upon the messages and stock value over the preceding day in [95]. Naive Bayes, decision trees and bootstrap aggregation (bagging) are used to learn the parameters for the proposed model. In particular, each message is classified into two *sentiment* classes "StrongBuy" and "StrongSell" and then *sentiment* prediction can be done with high accuracy and high recall. A trust value for each author is also considered: for each message, the author's *sentiment* is predicted and compared to the actual stock performance, thus if the author's post supported the actual performance, his trust value is increased. The classifier is then learned to predict whether the stock price of Apple, ExxonMobil and Starbucks will go up or down using all the features, including *sentiment* and trust vales, of the preceding day. The proposed model was able to make predictions with an accuracy ranging from 63% for ExxonMobil up to 81% for Apple.

In [96] the Multiple Kernel Learning (MKL) method is used to train a support vector machine (SVM) with an adaptively-weighted combination of kernels which embed different kinds of features. The resulting model is then applied to predict the stock prices for three Japanese Technology companies (Sharp, Panasonic and Sony) after the extraction of features

from financial time series data (exploiting technical analysis over price and volume) and from news and comments posted on the Engadget website (a web magazine with obsessive daily coverage of everything new in gadgets and consumer electronics).

Some technical indicators, such as the Rate of Change (ROC) and the Moving Average Convergence-Divergence (MACD), that are used to identify trends, and the BIAS indicator, that is used to catch overbought or oversold signals, are extracted from the time series of price and volumes of the three considered companies thus providing a first set of features for the MKL algorithm.

SentiWordNet [97] is used as the *sentiment analysis* tool to analyse the overall *sentiment* (objective, subjective, positive and negative) for each news and comment. The number of positive, negative and neutral news and comments for each target company, together with their frequency, is then considered to be the other set of features for the MKL algorithm.

The final step of the proposed methodology is to predict the stock price based on a MKL regression framework and evaluate the prediction results using *MAE*, *MAPE*, and *RMSE* evaluation measures.

The analysis performed from January 1, 2006 to August 15, 2008, over twelve shifting periods, each formed up by 8 months to train the model and 2 months for testing it, showed that the model exploiting MKL with all the considered features together outperforms the baseline methods, which are support vector regressions dependent on subsets of all the considered features. The accuracy of the MKL based model is on average 3 times higher than that of the other models.

Another interesting finding regards the clout of the MKL coefficients relative to the features of the *sentiment analysis* and the frequency of news and comments. In the case of Sharp, the coefficients relative to news and comments increased their influence abruptly between the training period starting in December 2006 and the training period starting in September 2007, which roughly corresponds to the sharp raise of Sharp stock price from December 2006 to March 2007. In the case of

Sony, since the number of news and comments is 6 times greater than the one of the other two companies, in all the MKL training periods the coefficients relative to the news and comments features account for more than 70 per cent of the total, thus showing that features extracted from sufficient news and comments may contribute more to stock price prediction than features extracted from historical prices and volumes.

Bollen et al in [90] investigated whether measurements of collective mood states derived from large scale Twitter feeds are correlated to the values of the Dow Jones Industrial Average (DJIA) over time.
The collections of daily tweets were feed into two different mood assessment tools, OpinionFinder and GPOMS, in order to generate a total of 7 public mood time series: a simple emotional polarity (positive or negative) together with other 6 different mood dimensions, namely Calm, Alert, Sure, Vital, Kind and Happy.
After having proved the effectiveness of the two *sentiment analysis* instrument in responding to significant socio-economic events such as the 2008 Presidential Election and Thanksgiving, the authors performed a Granger causality analysis of two different models, one based exclusively on past prices of the DJIA and the other based on past prices of the DJIA together with the 7 mood dimensions. The result showed that Calm dimension with lags ranging from 2 to 5 days is Granger causative of price movements of the DJIA. Hence, the Calm mood dimension has been proven to have predictive value with regards to the DJIA.
A Self-Organizing Fuzzy Neural Network (SOFNN) model was then used to predict DJIA values on day *t*, basing on a combination of DJIA values and mood values of the past *t-1*, *t-2*, and *t-3* days. The training of the model was performed on the period ranging from February 28, 2008 to November 28, 2008 and the final testing, performed on the period ranging from December 1 to December 19, 2008, yielded an impressive forecasting accuracy, a *Mean Absolute Percentage Error* (*MAPE*) of 1.83% and a direction accuracy of nearly 87% regarding the SOFNN

model with past values of the DJIA and the Calm mood dimension. The two baseline, that were the model with only past values of the DJIA and the model with past values of the DJIA together with the positive/negative *sentiment* values both yielded a MAPE of 1.94% and a direction accuracy of 73.3%.

A step further with respect of [90] is [91], where 476 million tweets from June 2009 to December 2009 were still analysed with a multi class classification, but this time only considering Calm, Happy, Alert and Kind mood dimensions. Furthermore, 4 different learning algorithms (i.e. Linear Regression, Logistic Regression, SVMs and SOFNN) were used to learn and exploit the actual correlation; however, SOFNN based model performed best among all other algorithms, giving nearly 76% of accuracy. Another headway with respect to [90] is the use of a k-fold sequential cross validation technique to measure accuracy so to train on all days up to a specific day and then test on the next 5 days. The authors also showed that the best correlation is between the happy and calm mood dimensions with the DJIA values.
A simple portfolio management strategy is also proposed, where buy decision is taken if the predicted stock value for the next day is a standard deviation less than the mean, while sell decision is taken when the predicted stock value is a standard deviation more than the actual adjusted value. The profit obtained using the proposed strategy was of 540 index points.

A platform different from message boards and discussion forums is embodied by Seeking Alpha, that is a popular financial social-media platform for investors to voice their opinions and exchange investment ideas. Authors on Seeking Alpha are required to reveal their identity as well as their positions on the discussed stocks and their opinions are submitted in the form of articles, which are then reviewed by a panel and are subject to editorial changes so to improve the quality of the published articles without interfering with the author's original opinion.

An evidence that *sentiment* revealed through Seeking Alpha has a larger and longer-lasting impact on stock *returns* than views expressed in more traditional media outlets, such as the Wall Street Journal, is given in [98]. In the attempt to examine whether social-media *sentiment* has a greater effect than the traditional media outlets, *sentiment* is extracted from both articles published on Seeking Alpha and from articles published in the Wall Street Journal print edition. For both sources the *sentiment* for company i on day t is set to be the average fraction of negative words across all articles published by that given source on company i on day t, thus obtaining the independent variables.
The regression of holding-period *returns* on the measure of traditional media *sentiment* and on measure of social media *sentiment* showed that when the holding period includes the day during which the media *sentiment* is computed, both types of *sentiment* are strongly correlated with stock market *returns*, whereas when the holding period does not include the day during which the *sentiment* is computed, only social media *sentiment* associates with stock *returns*.
Another relevant finding in [98] is that stock ownership has a moderate relevance on the impact of media *sentiment* on stock *returns* and this is shown for a subsample of firms for which more than 60% of the outstanding shares are held by retail investors. This evidence supports the fact that opinions revealed through Seeking Alpha have a greater potential to alter market prices when securities are primarily held by retail investors.

In [99] Chyan, Hsieh and Lengerich, from Stanford University, focused their attention in proposing a software agent which maximizes return on investment by trading the Dow Diamonds (DIA) ETF (The Dow Diamonds (DIA) is an exchange-traded fund that holds the 30 stocks that comprise the Dow Jones Industrial Average), exploiting *sentiment* of tweets and past values of the same ETF.

Sentiment analysis was performed over all messages posted on Twitter from June 12, 2009 to December 31, 2009 and the direction of change in market opening price, happened during the same period of time, was scored as a +1/-1 binary classification. As the major finding in [90] was that only calm mood dimension is useful for predicting the DJIA *returns*, only this specific mood dimension is considered in [99]. The tweets for each single day were scored using a bag-of-words method, where the calm score for that day were incremented by one for each occurrence of a word belonging to the calm dictionary.
A neural network was selected as the modelling tool and stated that the correlation between the DIA and the calm score is non-linear. The neural network was then trained to predict direction changes of the stock price for each day, a first time taking as input the sole time series of the previous 3 days of changes in DIA opening price and a second time taking as input the price time series together with the previous 3 days of Twitter calm score.
A policy search reinforcement learning agent takes as input the output of the neural network with the objective of buying and selling a certain amount of stocks so to reach the desired percentage of investment taking the optimal amount of risk. The reinforcement learning agent was trained on the period from July 20, 2009 to December 1, 2009 and tested from December 2, 2009 to December 30, 2009. Even though Twitter calm mood improved the prediction accuracy of the neural network by 5%, the agent profits more without the mood score.

At the time of writing there exists some *sentiment analysis* specialized firms that are exploiting all the current state of the art techniques in behavioural finance, data mining, machine learning and computational linguistic techniques to extract all the financial relevant knowledge from all the possible sources such as social networks, chat, forums, news site and corporate website [100][101].

Some trading platforms, among which the most resounding one is eToro [86], are even embedding technologies in order to allow investors to form social network ties between each other and to copy others' trades. Furthermore, it has been discovered that generally social trades outperform individual trades and that social influence plays a significant role in users' trades, especially in case of decisions during periods of uncertainty [102].

Not surprisingly, hedge funds are looking with deep interest to this growing body of new technologies and specialized firms, so they started to pull alongside traditional financial indicators, such as the P/E ratio, the more recent social media *sentiment analysis* [93]. For instance, the Derwent Capital Absolute Return Fund started its business in 2008 and since then it is providing investors and traders the opportunity to monitor global *sentiment* on stocks, currencies and commodities in real-time [85].

This brief introduction to this exciting and novel approach to stock market prediction is what motivated our work.

Chapter 3

Qualitative and exploratory analyses

The numerous critics [103][104][105] to the *Efficient Market Hypothesis* together with the reasonable amount of works (cfr. Paragraph 2.6 of Chapter 2) that have tried to demonstrate the predictive power, or more broadly the correlation, of social mood on the stock market form the two main hypotheses that support this work. It is worth noting that even the Nobel prize for the economy Paul Krugman criticized the Fama's famous *Efficient Market Hypothesis* thus supporting the behavioural finance thought that real investors are somehow irrational and subject to the behaviour of the crowds [106].

These financial views are helpful in taking the right perspective on this work and constitute the cue to the development of a methodology based on the fact that fundamentals data alone are insufficient to thoroughly explain daily *returns*; hence, a complementary reason should be explored in the mood of a vast number of people which can be considered to be representative also for the so-called noise traders which have a considerable effect on the stock prices volatility [107][108][109].

In this chapter we are going to give an overview of the preliminary analyses that have been carried out to approach the problem of predicting the stock market by using *sentiment* data.

3.1 Why Twitter?

The first decision we faced was the choice of a suitable data source for getting the so-called public mood about stocks and finance in general. In particular, we focused our attention on three main platforms, with different pros and cons:

- *Finanza Online Forum:* it was the first data source we analysed. It is a popular Italian online discussion site, where people can talk about many different topics related to the stock market. At first, we started to believe that it would have responded to our needs, because it is well structured (therefore, it is easy to retrieve all the posts related to a specific topic) and because its users are real traders that share their opinions and investment strategies. However, a deeper analysis revealed that the volumes of chatter were too low for our purposes. Moreover, the users show the tendency to get-off track in conversations. Finally, this choice would have limited the applicability of our approach to the Italian stock market.
- *Yahoo! Finance:* it is a popular website that provides financial news and information. "It has been the top financial news and research website in the United States since January 2008, with more than 37.5 million monthly unique visitors" [110]. It maintains a page for every stock and index in the worldwide markets. Furthermore, a lot of stock pages have a message board that is mainly used by users to read and post comments about financial news. The problem, once again, is that the messages are few with respect to those that are continuously published on social networks.
- *Twitter:* as we said in the previous chapter, it is the second online social networking service for number of users. That being said, it

is evident that it overcomes the problems associated with low volumes of chatter and it offers highly dynamic and diverse information. Another advantage in using Twitter instead of other sources like online forums is that such service exposes APIs to retrieve and filter the messages, even in real-time. Besides, Twitter and social networks in general are an immediate medium of information, that allow users to instantly share contents with their network of acquaintances. This is even truer now, with the pervasive diffusion of smartphones and mobile devices. Moreover, as explained in details in the previous chapter, Twitter characteristics make it quite close to human social networks. Finally, a specific peculiarity made it really interesting for our purposes: "Twitter usage noticeably spikes during disasters and other large events" [111]. It turns out that *Twitter buzz* is particularly useful to accurately capture the consequences of real world events on public mood and that this kind of information can theoretically be directly exploited to predict the impact of financial events on the stock market.

3.2 Twitter as a news medium

After choosing Twitter as data source, we tried to evaluate the importance of Twitter as a news medium. In particular, we focused on financial news and we tried to assess if crawling Twitter would provide or not an advantage in terms of time with respect to other sites. This was a fundamental aspect and the basis over which the study was built. In fact, if Twitter reported news with a significant delay with respect to other sources, it would be useless to predict future trends, while if it anticipated news sites, it would provide a direct advantage to investors intent on monitoring it. In order to investigate this point in detail, we conducted a one-week study, monitoring both Twitter and Google

Finance. We collected 135 news published on Google from March 12, 2012 to March 16, 2012. Specifically, we focused on articles that were related to a predefined list of specific brands belonging to the automotive and pharmaceutical sectors. The methodology adopted was the following: we started to retrieve news and their timestamps from Google as well as the timestamp at which the article was published on the original website. After that, we listed the most important keywords associated to each news and started to submit queries to Twitter through the search APIs in order to get the same piece of information. All the results were stored and manually examined to ensure that we got what we expected. After filtering out all the irrelevant messages and thus increasing the query precision to 100%, we compared the timestamp of the oldest tweet (i.e. the first to have been published) with the one that appeared on Google Finance. Finally, we computed the difference between the two timestamps and averaged them all. We obtained the following results:

- In the case of the Automotive Industry, on average, Google reports the news 4 minutes before Twitter.
- In the case of the Pharmaceutical Industry, on average, Google reports the news 2 minutes after Twitter.

It follows that using Twitter as a news medium provides neither advantages nor disadvantages in terms of time delay with respect to other sites. This is probably due to the fact that this type of information is more likely to be copied on social networks from other sites, rather than being generated for the first time on Twitter. In fact, we observed that news are mainly spread by automated profiles that catch information like the Google crawler does. Thus, the time required to find and propagate the news on Twitter and Google are equivalent. Finally, this time is so low that it is quite impossible to improve the performance.

Table 3.1 shows the top 5 posts authors of our data set in terms of frequency, i.e. the number of times we chose them for their timeliness in

reporting news. Simply from the names, it is not difficult to see that they are all automated tools, as said before.

Table 3.1: Posting frequency of the most active users on Twitter

Twitter account	Frequency
automatedtrader	8
CAC40feed	7
SeekingAlpha	5
wallstCS	4
thefinancepress	3

These profiles differ from the top 5 news site cited from Google, as shown in Table 3.2.

Table 3.2: Frequencies of appearances of news sites on Google

News site	Frequency
www.businessweek.com	17
www.4-traders.com	11
www.bloomberg.com	11
www.marketwatch.com	10
www.foxbusiness.com	8

All these sites also have a profile that automatically publishes the articles on Twitter, but with a delay greater than the time required by Google and other Twitter profiles to report the same news. This is probably due to a policy that privileges the users that visit directly the sites instead of Twitter.

Therefore, we were forced to conclude that Twitter cannot be exploited solely as a news medium and we made the assumption that Twitter should contain some latent information that impacts on the decisional processes of market makers and therefore on stock market prices. This conjecture drives the subsequent work, where the *Twitter buzz* will be used for predicting changes in stock prices.

3.3 Linear and nonlinear models

In order to achieve this goal, we decided to build a dataset comprising both financial data and online chatter and to try to build a model that could explain such dataset. Therefore, another decision we had to face was the choice between linear and nonlinear systems.
"In mathematics, a nonlinear system is one that does not satisfy the superposition principle, or one whose output is not directly proportional to its input; a linear system fulfils these conditions. In other words, a nonlinear system is any problem where the variables to be solved for cannot be written as a linear combination of independent components." [112]. Linear systems exhibit features and properties that are much simpler than the general nonlinear case [113], and for instance Bollen et al. [90] suggest that "the relation between public mood and stock market values is almost certainly non-linear". However, there are some serious flaws in Bollen's work that made his methodology not replicable in our study and encouraged us to explore more in depth the field of linear systems:

1. He collected all the tweets published from February 28 to December 19, 2008 with the only requirement that they contained statements that somehow express feelings, such as "I feel...", "I am", etc. On the contrary, we started from the reasonable idea that public mood must be measured by filtering in some way the

online chatter. For instance, if an abrupt change in Ford stock price occurs and if it is related to a correspondent change in the public mood, this must be searched in tweets containing at least the word "Ford" or even a financial statement of some sort. We assumed that posts like "I feel sick" or "I feel like I'm talking to myself if I don't get retweeted" are too generic and so there is no reason for them to have an impact on the stock market.

2. For the same reason as above, they were forced to perform a Granger causality analysis for the period of time between February 28 to November 3, 2008, thus excluding big socio-cultural events like the American presidential election and the Thanksgiving day because of the bias they inevitably introduce. This fact reinforced our conviction that a methodology of general applicability (i.e. independent from the period considered) must be founded on a further refinement of the input data that should be categorized according to their content.
3. Finally, his testing approach is not so rigorous. Firstly, as a general rule of thumb, training and testing sets must be big enough to be statistically significant. If this is the case for the training set (from February 28, 2008 to November 28, 2008), the same does not hold for the test period, that lasts from December 1 to December 19, 2008. This means that only 15 points (20 days minus weekends) are considered for testing. Moreover, he concludes that the calm dimension of public mood improves the prediction accuracy with respect to the case in which only past values of DJIA are used to train the Self-Organizing Fuzzy Neural Network. However, we decided to refuse this conclusion for a simple reason: an improvement of 13,4% on a set of 15 points means that the model considering both financial data and mood correctly predicts only two outcomes more than the benchmark, and this may be trivially due to chance.

Hence, we decided to adopt a different approach and to add complexity to the process of tweets collection and categorization. With regard to the model, we chose to start from the very basis and opted for linear systems.

3.4 Initial experiments

After having chosen the linear modelling approach, we started to collect some data and to conduct some experiments in order to have a better insight on the topic. In fact, we assumed that an obvious linear correlation between the *sentiment* and stock *returns*, if present, should be visible and detectable even with a manual examination. In any case, such an examination would have been useful also for inferring possible corrections to our hypotheses. Therefore, we decided to postpone the coding effort and to further investigate such point.

As a first experiment, for simplicity reasons, we focused on the Italian stock market (namely Borsa Italiana) and in particular on the stock related to Fiat, the Italian automobile manufacturer based in Turin. We considered Twitter as the unique data source, for the reasons explained in the first paragraph of the present chapter. By connecting to the streaming APIs, we collected all the tweets published between March 19, 2012 and March 20, 2012. Moreover, another crawler was employed to gather data about the stock related to Fiat from BorsaItaliana.it, with the granularity of the single contract.

Firstly, a pre-processing step was applied both to tweets and to stock prices. We discarded some tweets in order to retain only those published during the normal *trading session* of the Italian stock market, which is from 9 a.m. to 5:25 p.m., and those published during the five hours preceding the opening. A total of 535 posts remained, 251 of which for the first day and 284 for the second one. With regard to stock prices,

they were averaged in order to have a single value for each hour of the *trading session*.

Secondly, we examined all the tweets in order to perform a manual *sentiment analysis*. This is a quite hard task for an automated system as well as for a human, because there will always be a certain level of subjectivity, as shown in [114]. Therefore, we decided to independently analyse all the messages, assigning them a value of *sentiment*, and finally comparing the results. All the posts to which two different values had been assigned, were further analysed by both of us until we reached an agreement. The *sentiment* was categorized into positive, negative or neutral and a three-value scale was adopted:

- *Positive:* +1
- *Neutral:* 0
- *Negative:* -1

A brief glance to the dataset helped us to discover the most common patterns and to make some reasonable assumptions in order to ensure consistency during all the evaluation process:

1. Tweets reporting uncertainty should be categorized as negative, as uncertainty is expected to have a negative impact on stock market prices.
2. Tweets about commercials, ads and new Fiat products should be categorized as positive, because usually a user shares these contents only if he enjoys them.
3. Tweets reporting news should be categorized with the same *sentiment* of the news.
4. Tweets reporting generic or ambiguous news should be categorized as neutral.

Thirdly, we aggregated all the tweets on the same hourly base like we did for financial data, in order to be able to compare the two trends. For

every hour of the *trading session*, we computed the sum of the values assigned to all the tweets published in that hour, while for the period preceding the stock market opening the posts were aggregated all together in a unique time point labelled with 0. Figure 3.1 shows the overlap between the curve of the prices and the histogram of the *sentiment* during March 20, 2012.

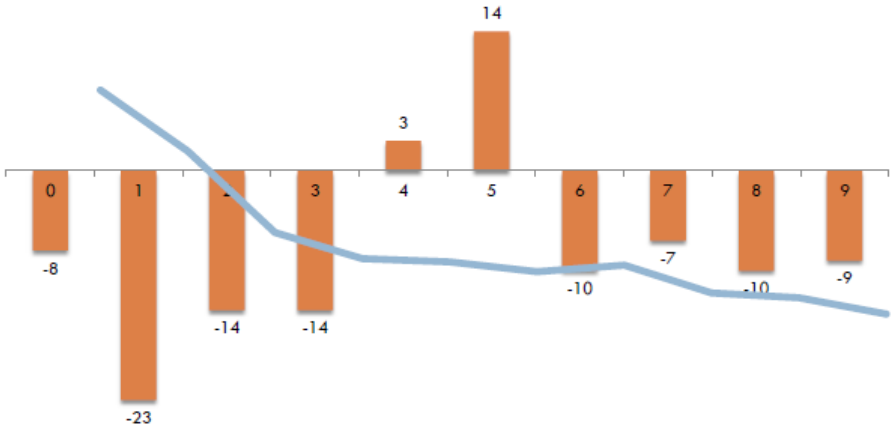

Figure 3.1: Overlap between the curve of the prices and the histogram of the sentiment during March 20, 2012

A value of -23 for the time point 1 means that the number of negative posts during that hour was equal to the number of positive tweets plus 23. The same holds for the other time points. A quite similar graph was obtained for March 19, 2012.

Thus we concluded that no obvious correlations between the *sentiment* and stock *returns* exist, even though an almost perfect assignment of *sentiment* values was performed. Hence, we tried to find some explanations that would have helped us to improve the theoretical foundations over which our work was being built:

1. The number of messages considered was too low to have a relevance of some sort. Thus, the aggregation on an hourly base seemed to be impractical to be used and we opted for a daily base

for future experiments. This choice was encouraged also by the fact that the mere difference between positive and negative tweets belonging to the time band 0 seemed to be predictive for the difference between the opening value of the day in exam and the closing value of the day prior to that. Of course, a two-day experiment has no statistical value, but we took it as a starting point for the development of more refined techniques.

2. Italian tweets are few with respect to the English ones and showed a negative bias for the period in exam. This may be due to the Fiat union policy and more in general to the sales trend of automotive companies in Italy.
3. Fiat is a multinational company active in many different countries and the Italian market affects its sales figures and stock prices minimally.

As a conclusion, we established to focus our attention on English tweets and to take as a reference not a single stock, but rather an index that incorporates many different companies in order to have bigger volumes of chatter.

Chapter 4

Methodology

With the preceding chapter we have justified our position and the reasons that brought us to choose Twitter as the main source for the gathering of posts, English as the main language and linear as the class of models to use.
In this chapter we will explain how we have approached the prediction of stock prices *returns* by making use of the social mood.
We will start by presenting the *brand model* with its hierarchy of *categories* that will form the core for the *sentiment analysis* tool, then we will go through the set of *sentiment variables* that we have defined and finally we will end the chapter describing the algorithm that we have developed for predicting *returns* by exploiting the *sentiment* gathered from Twitter, which can be considered as the climax of the entire work.

4.1 Modelling the social mood

The first problem that we are going to address is to define a proper model that, mapping the *domain knowledge*, drives the selection of relevant posts and refines the *sentiment evaluation* process. It is worth noting that, due to planning reasons, we opted for the separation between the *sentiment analysis* part, for which we decided to rely on an external tool, and all that has to be done before it. In particular, we are

focusing our attention in modelling the *brand* within a corporate domain by means of a top-down approach. The attention is towards those aspects that we believe to be the ones that will have a direct impact on the price movements of the *brand* related financial instruments. Hence, the proposed methodology differs from the usual *brand model* definition done for *Web Reputation* purposes. In fact, many systems, whose goal is to provide a *brand* attractiveness model, usually employ complex bottom-up techniques, which starting from the analysis of a large set of web posts attempt to extract the key aspects for determining the *brand reputation*.

Nevertheless, our methodology may be indirectly similar to the one employed by *Web Reputation* systems for what concerns the *non-financial* issues, since we also consider those posts that do not belong to any *financial category* but are somehow expressing their author's opinion on the *brand* under analysis.

Although we differentiate the *financially relevant statements* from those that are not, we will let the forecasting algorithm choose the best set of variables, according to what we are going to explain later on during this chapter.

What follows is a detailed description of the methodology whose final result will be the *brand model* that will support the *sentiment analysis* execution.

4.1.1 Model definition

The process of defining the model is a fundamental part over which all the subsequent components are built. Hence, it is important to conduct the work with a well-defined methodology in mind, in order to avoid wasteful re-implementations.

The first task is the choice between a bottom-up or a top-down approach. A bottom-up tool tries to extract directly from the data what

is important, while a top-down methodology requires a domain expert who is responsible for defining variables that are crucial for the analysis. We opted for the latter for various reasons. First, we decided to develop a specific financial tool, not a general-purpose one, by verticalizing and properly modifying the general *Web Reputation* analysis methodology. Therefore, starting from this clear context, there was no need to have expertise in various fields and so to adopt a bottom-up approach as we were moving in a predefined domain, that is, all that concerns or may affect the financial aspects of the *brand*. Second, we aimed to define a model that would be quite stable, requiring only minor adjustments from sector to sector.

We start by pointing out some of the terms that are fundamental for the definition of the model in which performing our financial analysis, that differs from a generic *Web Reputation* analysis as described in the previous paragraph. Here is the list of terms:

- *Brand:* the subject to be analysed. It can be a single public traded company or an index related to a specific sector, such as automotive or pharmaceutical.
- *Keyword:* we define *keyword* any expression whose presence in a sentence ensures that what is said is related to a specific concept (i.e. the *brand*). The set of *keywords* is used to retrieve from the Web all the posts that are related to the *brand*.
- *Category:* the set of characteristics and features of the *brand* that should be monitored singularly (e.g. sales, hires, etc.). Each *category* can theoretically be declined in several *subcategories*, according to the desired level of granularity, thus forming a taxonomy of concepts related to the *brand*.
- *Semantic Concept:* they are simple semantically related terms or expressions aimed to drive the categorization process; each

semantic concept is linked to a specific *category* by means of a "categorized as" relation.

The adopted methodology provides a series of steps that correspond to the design and implementation of all the elements reported above.
The first step is the definition of the tree of *categories* and the selection of a set of *semantic concepts* that would be exploited in the categorization phase.
Secondly, a list of *keywords* must be compiled and refined until good precision and recall are attained. After that, the crawling of an initial sample of tweets may be useful to assess the adequacy of the model. Finally, after a stable point is reached, the tool can start to collect tweets in real-time, to clean their content and to forward them to the external component that is in charge of performing the categorization and the *sentiment evaluation*.
Figure 4.1 summarizes the process we have explained.

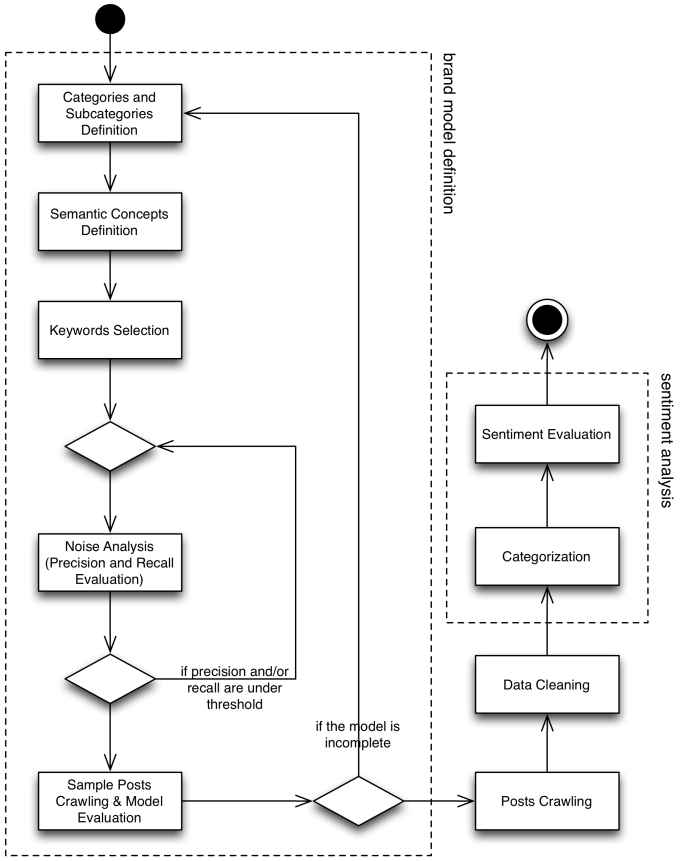

Figure 4.1: Activity diagram of the brand model definition and sentiment analysis

Keywords definition

As described above, the first step is the definition of a set of *keywords* that are exploited to filter the huge number of tweets that are constantly published on Twitter. In particular, the aim is to retrieve all and only those messages that concern some specific *brands*. For instance, the word "iPhone" is a perfect example of *keyword* for the *brand* "Apple", because "iPhone" is a registered trademark of Apple Inc. [115] and this implies that any sentence containing such word is surely related to Apple. In general, the definition of an effective set of *keywords* is a non-trivial task

that requires a deep knowledge of the specific application domain. It is a typical trade-off problem that requires a balance between precision and recall. Precision is "the fraction of retrieved instances that are relevant" [116], i.e. the number of collected messages that are actually related to the *brand* over the total number of retrieved posts. Recall is "the fraction of relevant instances that are retrieved", i.e. the number of collected tweets that are related to the *brand* over the total number of messages that are theoretically available on the data source. Recall grows as the number of *keywords* grows; therefore, the idea is to create a sufficiently large set of *keywords* without affecting precision. Going back to the previous example, "iPhone" ensures a precision of 100% but a low recall, because iPhone is just one of the Apple's products. Anyway, a good practice consists in choosing proper names as *keywords*, because they are more likely to be unique. However, a well-known problem of proper names is that they are subject to two kinds of ambiguity: weak and strong. A word is said to be weakly ambiguous if it is polysemous but its actual meaning can be deducted from the context. For instance, the word "Apple" is weakly ambiguous because we can distinguish between the fruit (e.g. "I ate an apple") and the multinational corporation (e.g. "I've just visited the new Apple store") quite easily. Moreover, proper names are also affected by strong ambiguity, that is, we cannot always derive from the context to which specific instance of a certain entity the author of a message refers to. As an example, if we collect all the tweets that contain the word "Gates", even if we are able to distinguish between the object and the surname, we cannot be sure that the author was referring to "Bill Gates" or to another person whose surname is "Gates". This problem can be obviously ignored if one instance (e.g. "Bill Gates") is far more popular than the others. As a conclusion, it is important to choose words and expressions that are not characterized by strong ambiguity. Conversely, weak ambiguity can be easily solved with the use of a word sense disambiguation algorithm. Therefore, we decided to adopt the following classes of *keywords*:

- *Brand name:* it is the simplest example of *keyword* and it is subject only to weak ambiguity.
- *Names of products:* as registered trademarks, they are neither weakly nor strongly ambiguous.
- *Key people:* e.g. CEO, chairman of the supervisory board, etc. As they can be strongly ambiguous, we opted to define *keywords* as couples *(name, surname)*, which are more likely to be unique. However, they usually appear in conjunction with the name of the *brand*, thus the problem of ambiguity is actually negligible.

The approach described is intrinsically top-down and general. Hence, it can be easily applied to almost all the sectors and to almost all public traded companies, with the only effort of searching for the most relevant products and key people's names.

Categories definition

The second step consists in defining a list of *categories* and *subcategories* and organizing them in a tree hierarchy. This kind of structure has the great advantage of avoiding the duplication of nodes, which in turn makes the categorization problem much simpler, given its single-label nature (i.e. every tweet is assigned to just one *category*), as opposed to a multi-label categorization, where *categories* can be partially overlapped. We decided to adopt a top-down approach, as opposed to a bottom-up approach in which an automated classifier is able to derive a partial or even a complete hierarchy of *categories* from a large enough sample of tweets. The reason behind this choice comes from the will to create a non-domain specific tree that could be employed in some different sectors with none or little changes. Obviously, it is impossible to map every aspect of the modelled reality. Hence, a further step of detection and analysis of trend topics is required in order to refine the model for each specific sector.

We report here the general-purpose hierarchy we have built following the design principles exposed above:

- *Employment:* it incorporates all that concerns organisational changes and working life dynamics. Its importance relies on the fact that big social phenomena, such as mass layoffs, may affect the *brand* perception. Its *subcategories* are:
  - *Hires*
  - *Layoffs*
- *Factory:* e.g. opening or closing of production plants, new facilities, etc. This kind of news can be a signal for future expansions in new markets or the beginning of economic problems.
- *Financial statement:* sales fall or growth, capital increases, reports and whatever financial aspect falls into this *category*. Since they are related to fundamentals data, they are expected to have a direct impact on the stock market. Its *subcategories* are:
  - *Earnings*
  - *Sales*
- *Innovation:* e.g. new products or services, research and development investments, the introduction of new production technologies, etc., i.e. everything which may offer a competitive advantage to the company under analysis.
- *Journalism:* it includes what popular journalists and news media have reported. Its importance relies on the fact that the press is able to influence public and investors' opinions.
  Its *subcategories* are:
  - *Journalists*
  - *Journals*
- *Management:* it refers to news and opinions about managers, management overhauls, key people's declarations, etc. It is one of

the most relevant aspects, to which special attention is given by experts and investors. Its *subcategory* is:

  - *Declarations*

- *Merger & Acquisition:* another fundamental point is to capture the *sentiment* of the chatter that is related to firms expansion policies, which may result in competitors ouster, expansion in new countries and markets, risk diversification, etc.
- *Other deals:* it collects everything that regards the company activity but does not fall into the other *categories*, such as agreements, collaborations, partnerships, etc.
- *Regulation:* laws, antitrust regulation measures and more in general whatever may force a company to change how it operates have an impact on the stock market, and thus should be monitored.
- *Strikes:* as we said about employment, strikes are a social phenomenon that may have a great echo around the world and so a negative impact on stocks prices.

Figure 4.2 shows the complete structure of the hierarchy.

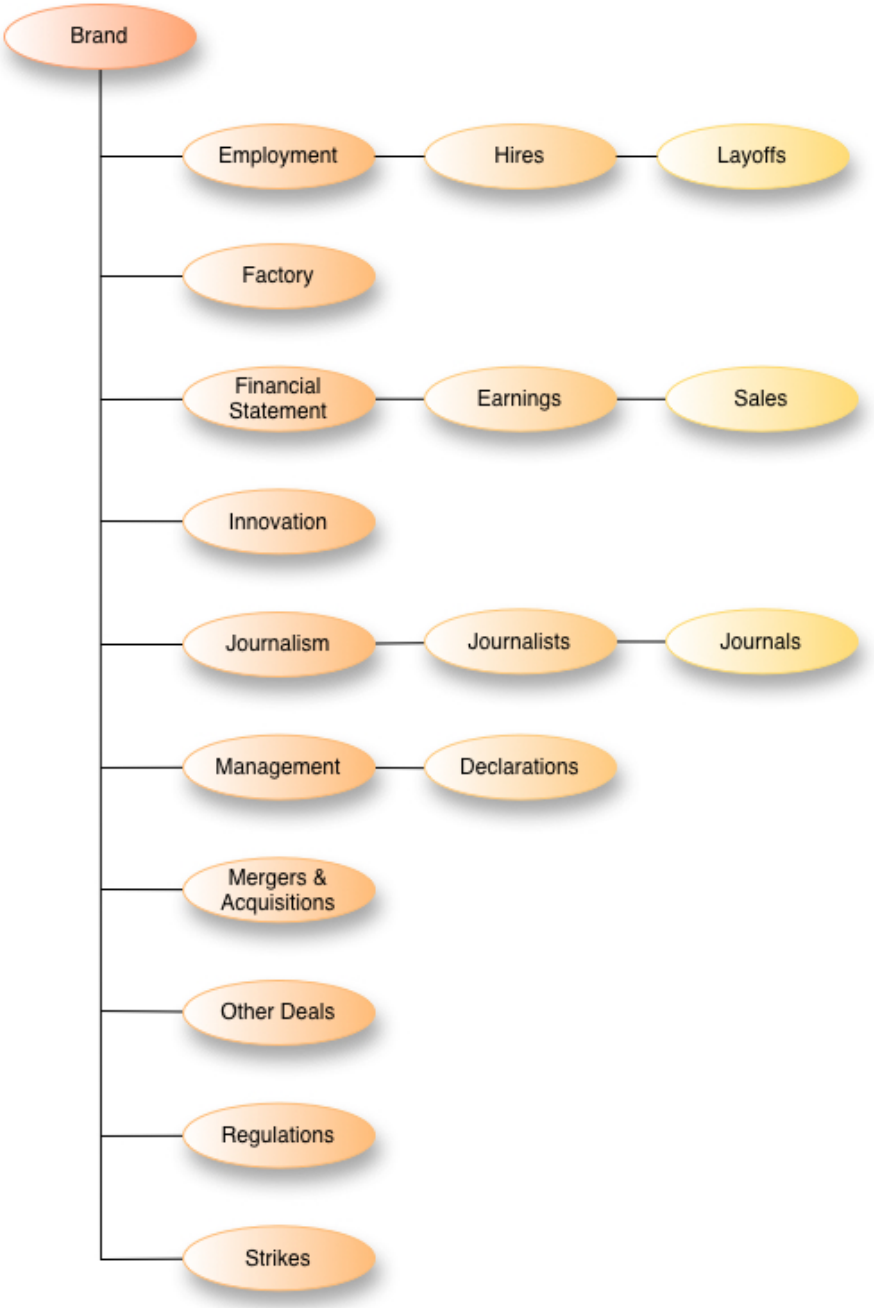

Figure 4.2: The financial categories hierarchy

For each *category*, a list of *semantic concepts* has been drawn up (cfr. Appendix). *Categories* and *semantic concepts* constitute the so-called *domain knowledge* that enriches and supports the external tool.
Moreover, all the previous *categories* fall into one huge class, called *Financial*, as opposed to *Non-Financial*. These two classes respond to the need of distinguishing between what is supposed to have a direct impact on the stock market (i.e. official news, declarations, etc.) and information that affects it indirectly. Hence, we decided to set up the generic *category Non-Financial*, that conveys what we called the *Twitter buzz*, which comprises a great variety of different kinds of information, that goes from opinions about products and advertisements to rumors, informal expression of feelings, etc. It is worth noting that *non-financial* content represents the vast majority of Twitter chatter and thus it should be taken into account as an integral part of the information source, as it may provide useful insights in better understanding the real perception of the *brand* by ordinary customers.

4.1.2 Implementation

What follows is a detailed description of the implementation of the various modules depicted in Figure 4.1.
The first component is a crawler that exploits the Twitter streaming APIs to collect tweets. Therefore, all the tweets containing one of the predefined *keywords* are gathered on a real-time basis.
The following step is about cleaning data. This is a necessary task in order to improve the data quality and standardize the content of tweets. In fact, even if online journals and some kinds of *UGC* are written in a correct and stylish manner, this does not hold in case of tweets for at least three reasons:

- *Twitter user base:* Twitter users are not (only) journalists or at least people that are used to write clearly and engagingly.

Moreover, this problem will probably worse over time as its user base is getting younger.

- *Immediacy:* the main goal of Twitter is to inform and allow people to write about what is currently happening, almost instantly. This need for immediacy is reflected in a writing style that is impetuous, rapid, informal and not thoughtful.
- *Narrowness:* a tweet must have a limited number of characters (less than 140) to be published and thus users are forced to condense their thoughts and to apply some tricks such as abbreviations, slang, shortened URLs etc.

The same applies, more in general, to different contexts because, as explained in the previous chapter, the sentence structure is strictly related to the specific community.

The cleaning process is made up of four stages, that have been designed specifically for tweets:

- *Filter:* it is the removal of elements that are typical on the Web but do not convey subjectivity nor any particular meaning (e.g. URLs) and the substitution of those that express *sentiment* (e.g. smileys) with an equivalent term.
- *Lingo:* it refers to words that have a meaning only in a specific context but can be safely removed. In case of Twitter they are *tags* (@), *hashtags* (#) and *retweet* marks (RT).
- *Slang:* slang expressions are substituted with a correspondent common word taken from the *Internet Slang Dictionary & Translator*.

After having cleaned up all the tweets, they are stored into a database along with the following information:

1. the timestamp of the tweet

2. the cleaned text of the tweet
3. the name of the *brand* to which the tweet refers to.

This data are then fed into the external *sentiment analysis* tool, namely SentiEngine [117].
Each post has to be assigned to a specific *category* by analysing the posts content and consequently exploiting the mapping between the *semantic concepts* and the *categories*. Finally, the categorized posts are further analysed in order to evaluate their polarity.
According to what has been exposed in the first chapter, SentiEngine main characteristics are:

- It is sentence level.
- It applies a lexical approach at semantic level.
- It assigns a monodimensional binary value for the *sentiment*.

This tool was preferred to other open source systems because it met all our requirements.
First, it takes care of semantically disambiguating the terms thus excluding the irrelevant posts. This layer of processing, in addition to the careful choice of a set of proper *keywords*, provides further confidence in the data quality.
Second, a monodimensional value for the *sentiment* allowed us to better estimate the values of precision. In fact, as explained in the previous chapters, this process is affected by a certain level of subjectivity and the problem gets even worse when the choice is between five or more values of a Likert scale.
Finally, the level of granularity to which we are interested is that one defined by the taxonomy of *categories*. *Sentiment* values are associated to a specific topic, that is, the subject expressed by the *category* to which the post has been assigned, no matter what specific feature the author cites in the text. Hence, we exploited the categorization and thus the

model to refine the *sentiment evaluation* process, as aforementioned. Moreover, because of the characters limit of a tweet, we do not expect that people would talk about many different features in the same post. It is worth noting that this approach is reasonable in this specific context, not in general.

For each *category*, the system will have to provide the following collection of values for the whole time span where the granularity is the single hour:

- *totPos:* the total number of positive tweets in a given hour.
- *totNeg:* the total number of negative tweets in a given hour.
- *cntNeu:* the total number of neutral tweets in a given hour.

As our initial assumption was to distinguish between *trading* and *non-trading session* and to aggregate the *sentiment* according to this classification, it follows that a further pre-processing stage is needed. At the end of this step the aggregated values serve as an input to an ad-hoc predictive algorithm that will use both financial and *sentiment* data.

Firstly, the prices of the selected index or stock (i.e. *brand*), that are crawled day by day with granularity of half an hour or finer, are filtered in order to retain only the values half an hour after the opening and half an hour before the closing. This is due to the fact that half an hour is considered to be the duration of the price setting interval for any given stock. These values are then labelled either as *day* for the ones corresponding to the *trading session* or *night* for the others. A *day* value contains as opening and closing prices the opening and the closing prices of the corresponding *trading session*, while a *night* value contains as the opening price the closing price of the previous *trading session* and the opening price of the subsequent *trading session* as the closing price.

Secondly, the *sentiment variables* are aggregated along three different dimensions:

- *Financial:* if a single stock has been selected for the analysis, there is no need for aggregating along this dimension. Otherwise, *sentiment variables* associated to each index component should be summed up.
- *Temporal:* as said before, the *sentiment* must be aggregated in a proper way in order to be consistent with financial data. All the tweets posted during the *non-trading session* must be grouped; the same holds for the *trading session*. Weekends are a particular case of *non-trading session* that lasts for more than a day.
- *Category:* as the volumes of chatter of each single *financial category* are too low while considering a one day period, they are aggregated all together. Anyway, the categorization process still retains its importance, because it allows to distinguish between financially relevant and *buzz*. Moreover, future developments may assess the importance of some specific *categories* over the others. Such an evaluation would only require to discard the least important ones before performing the aggregation.

This method allows the creation of a temporally continuous dataset of stocks prices and the corresponding *sentiment*.
Thirdly, the so-aggregated *sentiment variables*, indicated in the following as *POS-(NON)-FINANCIAL*, *NEG-(NON)-FINANCIAL* and *NEU-(NON)-FINANCIAL*, are used to compute other variables. They are:

- *POS:* the sum of *POS-FINANCIAL* and *POS-NON-FINANCIAL.*
- *NEG:* the sum of *NEG-FINANCIAL* and *NEG-NON-FINANCIAL.*
- *NEU:* the sum of *NEU-FINANCIAL* and *NEU-NON-FINANCIAL.*

- *TOT-NON-FINANCIAL-SENT:* the sum of *POS-NON-FINANCIAL* and *NEG-NON-FINANCIAL.*
- *TOT-FINANCIAL-SENT:* the sum of *POS-FINANCIAL* and *NEG-FINANCIAL.*
- *TOT-SENT:* the sum of *TOT-NON-FINANCIAL-SENT* and *TOT-FINANCIAL-SENT.*
- *POSMINUSNEG:* the difference between *POS* and *NEG.*
- *POSMINUSNEG-NON-FINANCIAL-REL:* the difference between *POS-NON-FINANCIAL* and *NEG-NON-FINANCIAL* over *TOT-NON-FINANCIAL-SENT.*
- *POSMINUSNEG-FINANCIAL-REL:* the difference between *POS-FINANCIAL* and *NEG-FINANCIAL* over *TOT-FINANCIAL-SENT.*
- *POSMINUSNEG-REL:* the difference between *POS* and *NEG* over *TOT-SENT.*
- *SENT-REL-NON-FINANCIAL:* the difference between *POSMINUSNEG-NON-FINANCIAL-REL* and the average of *POSMINUSNEG-NON-FINANCIAL-REL* computed over a sample period. The aim of this operation is to depolarize the values, i.e. remove any positive or negative bias.
- *SENT-REL-FINANCIAL:* the difference between *POSMINUSNEG-FINANCIAL-REL* and the average of *POSMINUSNEG-FINANCIAL-REL.*
- *SENT-REL:* the difference between *POSMINUSNEG-REL* and the average of *POSMINUSNEG-REL.*

All these variables are supposed to be useful to model and thus predict variations in stocks prices. In fact, we assumed that no general model can be conceived; therefore, we started to design a system that will

constantly and automatically change the model in a proper way, exploiting past data to adapt itself to changes.

The following paragraph will present an algorithm that aims to prove our claims.

4.2 Predicting the stock market

As seen in Paragraph 2.6 of Chapter 2, where some state of the art methodologies were considered, various techniques have been proposed to exploit the alleged predictive power of the *sentiment* on *returns*, recommending different kinds of linear as well as non-linear models. Moreover, there are plenty of prediction algorithms that manipulate time series, both *black-box* and *white-box*. However, we decided to distance ourselves from these kinds of models and to develop an ad-hoc algorithm for various reasons. First, *black-box* models (e.g. neural networks) or *grey-box* ones do not allow to easily understand what is actually happening inside the system. Therefore, if the dynamic behaviour of the system is not that desired, it is quite impossible to understand the causes and make the right corrections in order to improve the predictive performance. On the other side, *white-box* models (e.g. *autoregressive–moving-average*) require a deep knowledge of the modelled phenomena, because they are often based on physical equations (or at least take as inputs some parameters whose meaning can be clearly deduced from the observation of the modelled reality) and thus on a top-down approach. But this is not the case for our work, as there are no underlying physical laws nor clearly visible parameters that influence stocks trends. It follows that the most reasonable way of proceeding was to take a data-driven approach and to try to design an algorithm that would infer some latent relations from these data in a manner that we could control and possibly modify.

4.2.1 The predictive algorithm

The proposed methodology has at its core the identification of different linear models based on a combination of *sentiment* and *financial variables*. What follows is a detailed description of the proposed methodology. The overall conceptual architecture is illustrated in Figure 4.3, where each component is labelled with a number.

The *sentiment variables* are provided by the aforementioned *sentiment analysis* component (number 3 in Figure 4.3) which, after having collected a series of Twitter posts (by mean of the component number 2 in Figure 4.3), classifies them according to different *categories* and assigns them a *sentiment* value, as described in the previous paragraph. The *financial variables* are provided by component number 1 of Figure 4.3, which downloads the prices of the selected *brand* from *Yahoo! Finance* and then computes the corresponding *returns*. *Returns*, instead of prices, of assets will be considered since this is what is done in most financial studies. [118] gives two main reasons for using *returns*. Firstly, for average investors, *return of an asset* is a complete and scale-free summary of the investment opportunity and secondly *return* series are easier to handle than price series because they have more attractive statistical properties.

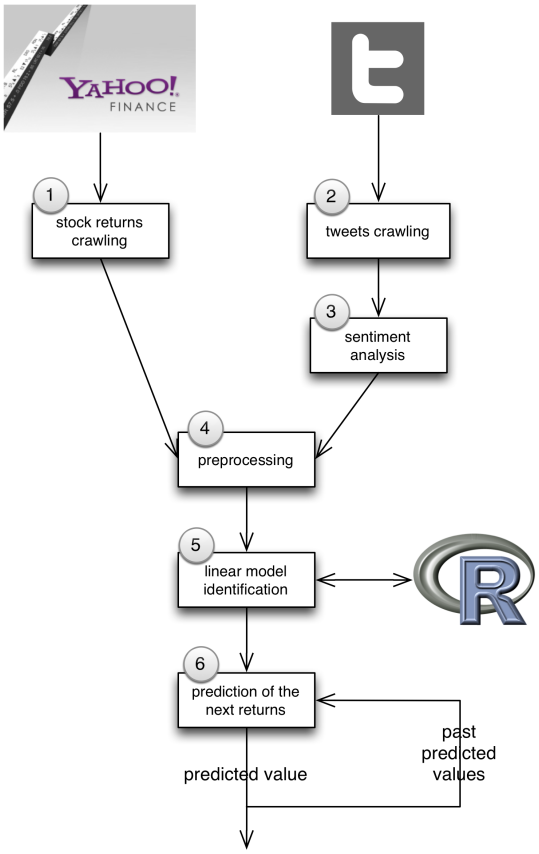

Figure 4.3: The overall conceptual architecture

Since there are several definitions of an *asset return*, in this work *simple net return* or *simple return* R_t are considered.

Let P_t be the price of an asset at time index t and assuming that the asset pays no dividends and that it is held for k periods, from time instant t-k to time instant t, the *simple return* R_t is defined as in Formula 4.1:

$$R_t = \frac{P_t - P_{t-k}}{P_{t-k}} \tag{4.1}$$

It is worth noting that, as stated before, two kinds of time periods have been distinguished, namely *day* and *night*. The former is related to the

time frame in which the stock market is open and during which the prices of each stocks are subjective to variations due to trading (i.e. the *trading session*). The latter is related to the *time frame* during which the stock market is closed so no trading activity can affect the stock prices (i.e. the *non-trading session*). The previous distinction is essential for the aggregation of the *sentiment* and for the computation of stocks *returns*.

Then a pre-processing stage (number 4 in Figure 4.3) groups the *financial variables* together with the *sentiment variables* in such a way that *returns* and *sentiment* values are placed in the correct time point with respect to the previously explained time period. Moreover, this phase is also in charge of handling the difference in terms of time zone between the tweets and the selected *brand*. In fact, as Twitter APIs retrieve a timestamp that is consistent with the crawler time zone (i.e. the Italian one), it must be reconciled with the timestamp associated to financial data (i.e. the New York one).

The linear model identification (number 5 in Figure 4.3) and the prediction of the next *return* (number 6 in Figure 4.3) form the core of the system and will be described later on.

Figure 4.4 illustrates the high-level functioning of the predictive algorithm.

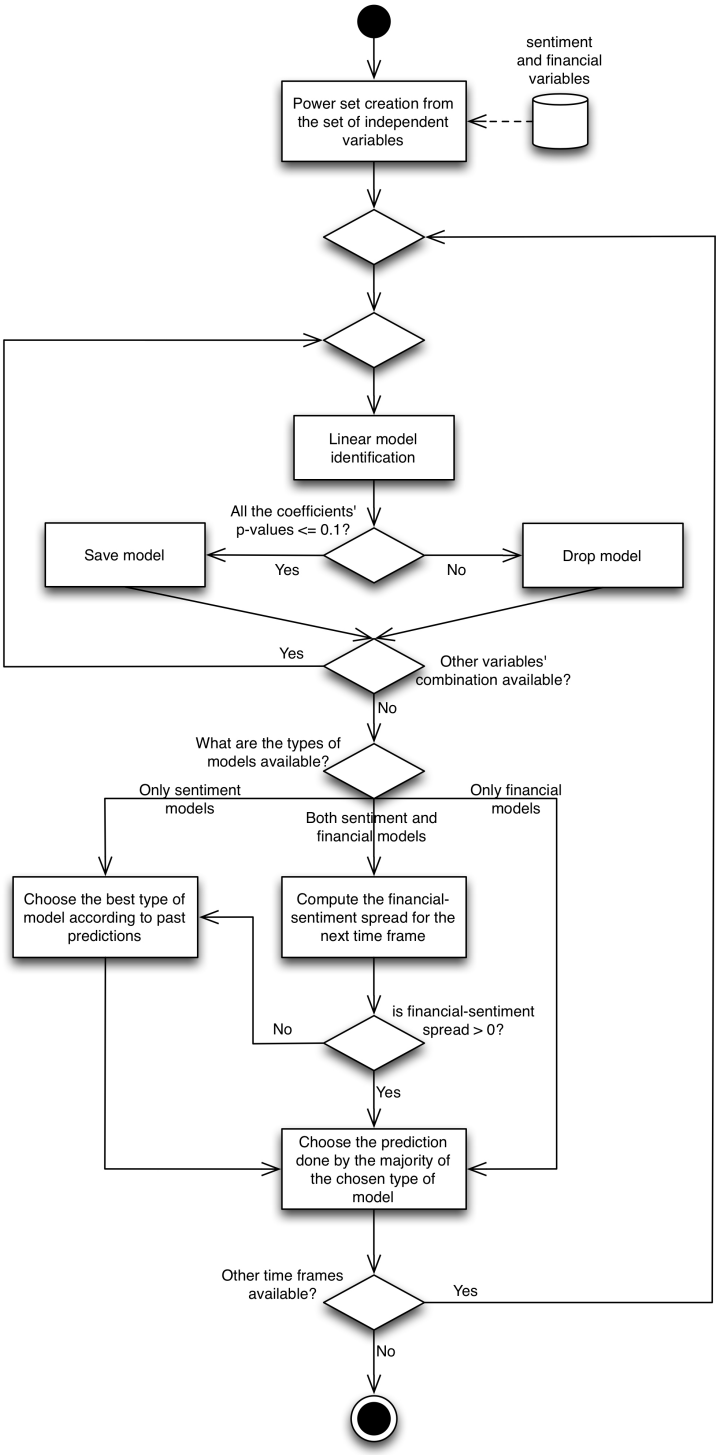

Figure 4.4: Activity diagram of the core part of the algorithm

A fundamental preliminary step is the creation of the power set of all the available variables. This step is where firstly takes shape the idea of an adaptive model. As a matter of fact, as it will be described in detail in the following chapter, we decided to merge the development of a predictive algorithm with the testing phase. This is due to the fact that, since the system as a whole cannot be conceived as an abstract entity, but rather as an integral part of the reality, as it is in charge of predicting real world outcomes, all the hypothesis made during the path must be carefully tested against the real world phenomena in order to prove their plausibility. Hence, as mentioned in the previous paragraph, we postulated that no general model can be conceived in order to explain the stock market phenomena and that for different *time frames* there could be different combinations of variables that best fit the data. In order to prove our belief, we built up a system that is able to test all the possible combinations of variables for a given period. As it was impossible to find a model for the whole period in analysis (cfr. Paragraph 5.7 of Chapter 5), we definitely concluded that the only viable way was to derive for each *time frame* a list of possible models and to teach the system how to identify the best one from time to time. It follows that the power set of the variables is an indispensable input that must be provided to the algorithm.
After that, the algorithm starts to analyse the first available *time frame* (namely *TF*) of a given *time frame width* (namely *TFW*), that is a set of adjacent time points to be considered in order to perform the identification of a model. Because of complexity reasons, Figure 4.4 refers to the case in which only one *TFW* is employed. The more general case that exploits a series of different *TFWs* will be described in detail hereinafter.

Subsequently the linear regression model identification is performed for each combination of variables, using the *lm* command of the *R statistical framework*, which computes the coefficient and the *p-value* for each

single variable that has been considered to build the model. The underlying idea is that the models that better explain a certain set of recent past values are more likely to correctly predict the next future outcome. Hence, all the models for which all the *p-values* of their variables are under a fixed threshold of 10% are kept for the forthcoming analysis. This choice is due to the fact that *p-values* (aka significance levels) represent the confidence that one can have in the estimated values being correct. In other words, *p-value* is a percentage that tells how likely it is that the coefficient associated to the independent variable emerged by chance instead of describing a real relationship. Hence, we decided to follow a common practice according to which variables with a *p-value* of less than 0.1 are considered as significant, that is, we retained as valid those models for which there is a 90% chance that the relationship is real and a 10% chance that the relationship emerged randomly.

The saved models are then classified according to the type of variables they use, i.e. if a model contains solely *financial variables* is considered to be a *financial model*, whereas it is considered to be a *sentiment model*. It is worth nothing that as *financial variables* we selected the pair made up by the *1-lagged* and *2-lagged returns*. The reason behind this decision was that *financial models* are purely autoregressive models, which are known to be useful in the analysis of financial time series, while *sentiment models* have an unknown usefulness until tried. Furthermore, *financial models* are supposed to behave in a quite different manner with respect to the *sentiment* ones, because the data series they use are diverse in nature: the former ones manipulate only small numbers that represent percentages, the latter ones a wide variety of data with different meanings. This hypothesis was confirmed by a subsequent analysis of the data (cfr. Paragraph 5.7 of Chapter 5), which showed that when *financial models* occur in conjunction with *sentiment models* they

often provide prediction values that differ in sign and thus are extremely diverse for our purposes.

Sentiment models are further classified according to their number of variables, thus obtaining $sentiment_1$, $sentiment_2$, and so forth till $sentiment_n$, where n is the number of the available variables. Once again, we needed a way to make a distinction among *sentiment models*, which may act quite differently from each other in terms of predictive power. At first, we hypothesized to treat each model as a single class in order to monitor its performance. However, a brief glance at the data hinted that such an approach would have been inappropriate (cfr. Paragraph 5.7 of Chapter 5). In fact, the number of occurrences of a lot of models was too low to infer anything. Therefore, we opted for decreasing the level of granularity and to aggregate models with the same number of variables. In fact, a general rule of thumb is to base the number of variables to be used on the number of available observations, i.e. the *TFW* [119]. As the algorithm exploits more than one *TFW*, it follows that the most reasonable number of variables must be chosen from time to time.
The key aspect of the algorithm is that it is adaptive, that is, it derives knowledge from its past performance and exploits this kind of information to adapt its behaviour and thereby hopefully improving the results. The adaptivity is embedded at two different levels: the lower level is the choice of the best model (between *financial* and *sentiment* or among all the available classes of *sentiment models*), given a certain *TFW*, the higher one is the choice of the best *TFW*. The underlying idea is that there is a certain degree of correlation in the same series of predictions and that they may provide some interesting and previously unknown cues. Thus, the history of predictions may be usefully employed as a further information source.
Therefore, at lower level, two indicators are computed in order to measure the quality of different classes of models. This is due to the fact that also here we can conduct the analysis at different levels and it seems

unreasonable to keep the same indicator for both cases. Moreover, the first indicator must be targeted on a pairwise comparison, while the second one should address the problem of choosing between many classes of models.

The first indicator is the *financial-sentiment spread* that drives the choice between *sentiment* and *financial models*. It is used to keep memory of the relative performance of the former ones with respect to the latter ones, giving an initial advantage to *financial models* by initially setting it to 1. As anticipated before, this is done mainly because *sentiment models* have to prove their superiority by making more correct predictions with respect to the *financial* ones. In fact, we assumed that *autoregressive* (*AR*) and *autoregressive-moving-average* (*ARMA*) models based solely on *returns* can provide some hints to their prediction and thus perform well in general, as shown in [120]. Hence, initially, the algorithm trusts *financial models* more than the *sentiment* ones, but it must be able to change its mind in useful time in order to respond to unexpected changes, that is, it must be reactive.

As stated before, this measure intervenes if both *sentiment* and *financial models* for a given *TF* exist, i.e. if they both meet the *p-values* constraint. In such cases, the *financial-sentiment spread* is incremented if the *correctness* of the *financial models* is greater than that of the *sentiment models* or decremented if the *correctness* of the *financial models* is lower than that of the *sentiment models*.

Now we have to clarify the concept of *correctness* that retains its value both at the higher (*financial* vs. *sentiment models* class) and the lower level (classes of *sentiment models*) of analysis. For each category, the value of *correctness* (namely *tm.perc*) measures the percentage of correctly predicted *returns* (i.e. predicted *returns* with the same sign of the real ones) that a specific class of models made in a certain *TF*.

A more detailed insight of the *correctness* measure is provided by the following pseudocode, where *TF* is a given *time frame*:

```
1. for every type of model tm in TF:
2.     tm.cpsum = 0
3.     tm.n = 0
4.     for every model m of type tm:
5.           if sign(m.predictedReturn) == sign(realReturn):
6.                 tm.cpsum += 1
7.           tm.n += 1
8.     tm.perc = tm.cpsum / tm.n
9.     tm.expMovingAverage = expMovingAverage(tm.pastPerc)
```

where *m.predictedReturn* is the *return* predicted by the model *m* for the time point following the last time point of the current *TF*, *realReturn* is the real *return* for the same time point and *expMovingAverage* is a function that will be explained later.

It should be emphasized that the *correctness* measure is based only on the sign of the predicted *return* because of a specific reason. We are interested into getting the correct prediction, i.e. a value with the same sign of the next future outcome, because this is the only information needed in order to decide when to buy or sell the stock, and the only one useful. Hence, in this sense, a prediction with the same sign of the real *return* but quite different in absolute value is more correct than one which is very close to the real *return* but differs in sign. It follows that whatever quality measure is chosen, it must take into account this fact and privilege the former prediction with respect to the latter one.

A pseudocode is again provided to better explain the *financial-sentiment spread* computation, where γ is a correcting factor between zero and one, and *abs()* is the absolute value function:

```
1. if "sentiment" and "financial" are both present in TF:
```

```
2.    if sentiment.perc > financial.perc:
3.          spreadFinSent = γ*spreadFinSent -
abs(realReturn*100)
4.    else if sentiment.perc < financial.perc:
5.          spreadFinSent = γ*spreadFinSent +
abs(realReturn*100)
```

If the *financial-sentiment spread* until *TF-1* is negative, i.e. *sentiment models* have outperformed the *financial* ones, the selection of the class of models is made in the same way as if only *sentiment models* were present in *TF*. Otherwise, *financial models* are taken as reference for the next prediction. The real *return* has been considered a good measure for updating the *financial-sentiment spread* because it takes into account the fact that wrong decisions have an importance that is proportional to their relative *return*. Therefore, it seemed appropriate to take as reference the cost of either a correct or a wrong prediction, i.e. the amount of money it would have allowed to earn or lose in principle.

The parameter γ has been introduced as a *discount factor*, i.e. in order to assign the correct weight to the history of predictions. For instance, if *financial models* performed much better than the *sentiment models* in the past and γ is too high, it would be quite impossible for the latter to invert the trend in the future, that is, the algorithm would not be reactive to changes. Likewise, if it is too low, it means that only the recent behaviour matters. Of course, as our approach was admittedly data-driven, we opted for not fixing it basing on a theoretical reasoning but rather we estimated it by the use of training data (cfr. Paragraph 5.8 of Chapter 5).

To further explain the effect of *discounting*, as it is going to be used again in the following, we provide the general Formula 4.2 that shows this effect where x_i is the *indicator* at time i, d is the *discount factor*, and

Δ_i is the *additive component* at time *i*, that is the new information for the *indicator* computation.

$$x_n = d^n x_0 + d^{n-1}\Delta_0 + \cdots + d\Delta_{n-2} + \Delta_{n-1} \qquad (4.2)$$

In the previous case, the *financial-sentiment spread* is taken as the x, the parameter γ is the d, and finally the absolute value of the real *return* is considered as the *additive component* (Δ_i) at each time step.

The second indicator is an *exponentially weighted moving average* (*EWMA*), which is a well-known type of "infinite impulse response filter that applies weighting factors which decrease exponentially" [121] that has been successfully employed in many applications, ranging from air defence radar to trading the Chicago pork belly market [122]. It is calculated for the series *pastPerc*, that is, the collection of all the past values of *tm.perc* (i.e. $perc_0$, $perc_1$, ..., $perc_{TF-1}$) from the first *time frame* (namely 1) in which *tm* was present to the last *time frame* (in which *tm* was present) before the one under analysis (namely *TF-1*). The *EWMA* has been chosen in order to give relatively more weight to recent *tm.perc* than to older ones and it is computed as follows, where α is the smoothing constant that is responsible for the speed at which the older *tm.perc* are smoothed [118]:

$$s_1 = tm.perc_0$$

$$s_{TF} = \alpha \cdot tm.perc_{TF-1} + (1-\alpha) \cdot s_{TF-2}$$

Once again, the parameter α is not fixed but it has to be estimated by the use of training data (cfr. Paragraph 5.8 of Chapter 5).

If only *sentiment models* are available in *TF*, the class $sentiment_i$ that has the maximum value of *EWMA* among the classes of *sentiment models* is then chosen to be the one to follow. Here we decided to

substitute the real *return* with *tm.perc* because the level of detail is higher and because a class of models is conceived as a whole, which means that profits are gained in a certain *TF* only if the majority of models inside it makes the correct prediction. Hence, the class with the best components, that is, the class that made the higher percentage of correct forecasts in the past is supposed to be likely to make a correct prediction in the near future and thus it is preferred to the others. In this sense, the algorithm integrates both a financial and a statistical measure of performance.

Finally, a brief testing (cfr. Paragraph 5.7 of Chapter 5) supported the idea of using two different measures for the two levels of analysis. Indeed, in addition to the aforementioned fact that they are different with respect to the number of classes of models participating to the evaluation, it is worth noting that some classes of *sentiment models* are rarely retained as valid, while this does not hold for *financial models*, which are more often present.

Once a single class of models has been chosen, the sign of the *return* predicted by the majority of the models within that class is then selected to be the predicted trend. This is because at this lowest level of granularity we cannot exploit any other source of information and we are forced to rely on the insight that the majority of a pre-selected list of good models will provide good forecasts. If no models are found to be relevant, the system does not yield any prediction.

The process just mentioned can be repeated on any number of *time frames* thus creating a series of predicted *returns* that can be compared with the real ones, as shown in Figure 4.5, in order to simulate an investment scenario.

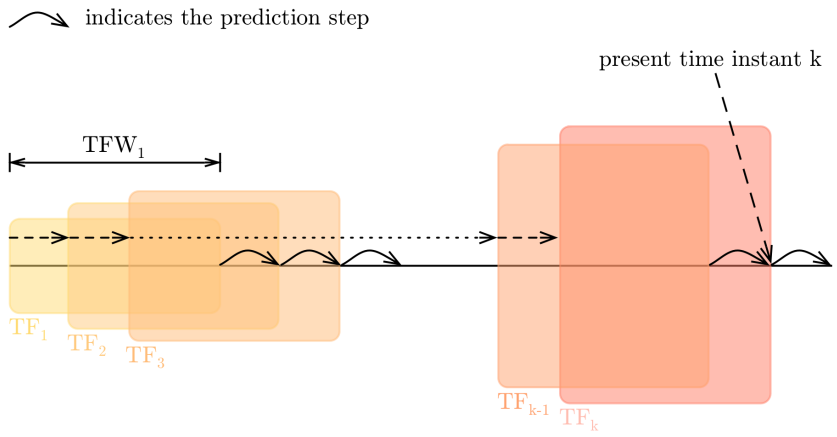

Figure 4.5: Scheme that depicts the prediction series resulting from the k times reiteration of the prediction part

The process displayed in Figure 4.5 can be started at an arbitrary time point in the past and can be iterated until it reaches the present time point, thus providing the one step ahead prediction. It is worth noting that, as stated before, this process requires the specification of a *TFW*.
Hence, at higher level, the same process can be performed with different *TFWs*, originating different *prediction series* that can be compared one another. It follows that different *TFWs* can be used in order to predict the real outcomes at different time points, according to which is the best one at any iteration. In fact, as we cannot know a priori which is the best *TFW*, the best thing to do is to include for example all the *TFWs* ranging from 20 time points (i.e. approximately one trading month) to 40 time points, rather than fixing it arbitrarily. Moreover, some *TFWs* may not produce any relevant model for a specific time point. Therefore, the inclusion of a sufficient number of *TFWs* helps to overcome such problem so to have one prediction value for every time point.
In order to make the comparison, a *quality indicator* (namely *QI)* associated to each *prediction series* has been thought and its computation is shown by the following pseudocode, where β is a correcting factor

between zero and one, *sign()* is the sign function and *abs()* is the absolute function:

```
1. given a prediction series PS with at least one element
2. QI = 0
3. for each TF in PS
4.    if not exists predictedReturn:
5.          QI = β*QI
6.    else:
4.          if sign(predictedReturn) == sign(realReturn):
5.                QI = β*QI + abs(realReturn*100)
6.          else:
7.                QI = β*QI - abs(realReturn*100)
```

The idea for updating the *quality indicator* is basically the same as that of the *spread*. In fact, this level of analysis is higher than that of *financial-sentiment spread* and thus it seems reasonable to keep the same high-level performance measure based on the investment scenario following Formula 4.2. Likewise, β retains the same meaning of γ but can obviously assume a different value, as the subjects of the comparison differ from one another. Hence *QI* is the *indicator* at time *i*, β is the *discount factor*, and the absolute value of the real *return* is the *additive component* at time *i*.

However, there is a simple but important difference: as most of the *TFWs* originate *prediction series* containing values for the majority of the time points (cfr. Paragraph 5.7 of Chapter 5), we can state that in this sense their overall behaviour is quite similar. Nevertheless, there are some particular phenomena that should be captured in order to further discount their quality measure.

Basically, we assumed that if there are some missing predictions within a *prediction series* originated by a given *TFW*, that is, no relevant models were found, it means that the usefulness of the *prediction series* created with that *TFW* for explaining the financial trend in that period has decreased. Hence, there is no reason for having so much trust in it when it will reappear in the future and its *QI* must be adjusted accordingly. Therefore, as shown by the previous pseudocode, *QI* is multiplied by β at each step, independently from the presence of the prediction value. Moreover, this is another way to increase the reactivity of the algorithm that is, its ability to rapidly discard those *TFWs* whose *prediction series* behaviour is quite unstable.

In order to better clarify the comparison between different *prediction series*, Figure 4.6 depicts two *prediction series*, one originated by the use of TFW_1 and the other originated by the use of TFW_n where the prediction done by TF_3 with a width equal to TFW_n is chosen due to the fact that the *prediction series* resulting from the use of TFW_n has a higher *QI* with respect to the *prediction series* resulting from the use of TFW_1.

It must be clear that an arbitrary number of *prediction series* can be simultaneously compared in order to select the *TFW* that has originated the *prediction series* with the highest *QI*. Finally, the chosen prediction is the one made by the majority of the models belonging to the best class, as explained before.

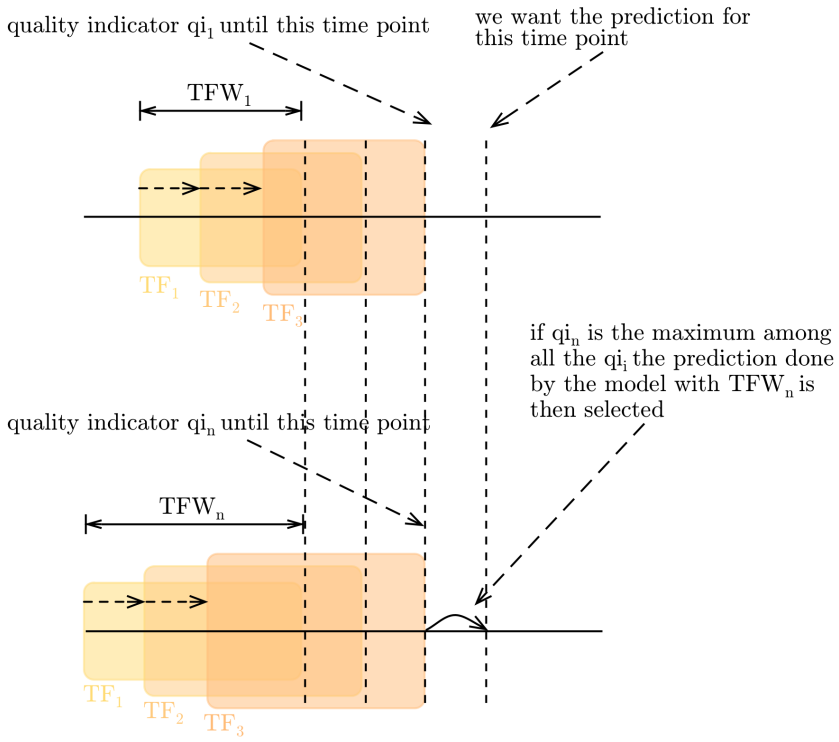

Figure 4.6: The comparison between different prediction series

4.2.2 Complexity

Assuming that the chosen stock market (for example the NYSE) opens at 9:30 and closes at 16:00, we open the *day* position at 10:00 and we close that same position at 15:30, as we said above. At the same time we open the *night* position, which will be closed at 10:00 of the first subsequent *trading session*. This originates a cyclical trading process. The posts crawling phase regarding the *day* goes from 9:30 to 15:00 of a given trading day, while that regarding the *night* goes from 16:00 of a given trading day to 9:30 of the first following trading day as depicted in Figure 4.7.

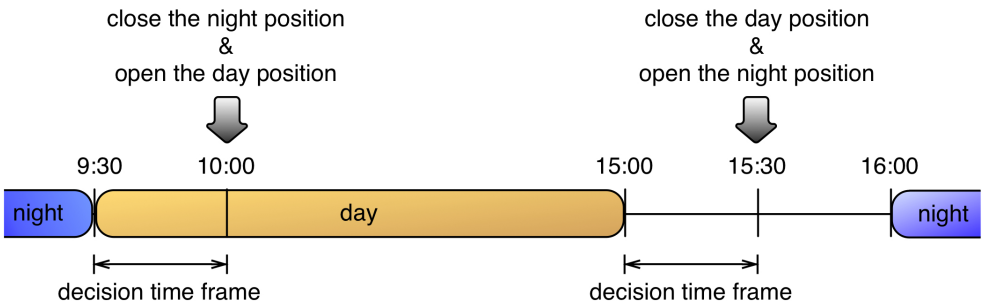

Figure 4.7: Timeline that represents the two posts groups that are gathered from Twitter assuming that the chosen stock market opens at 9:30 and closes at 16:00

That have been said, it is clear that 30 minutes is the time span during which we can perform the analysis of the last *TF* (i.e. the one that ends with the last session) and make the prediction that drives the decision on either opening a long position or a short position. In order to avoid confusion, let's focus our attention on the *trading session* and let's imagine we are at 9:30 of a given trading day: assuming that the *sentiment analysis* is performed in real time, the system has 30 minutes to either examine the past performance, that may be stored in a database and thus be readily available, and to analyse the last *TF*, that is the one ending with the last *night* also including the prediction of the forthcoming *day return*.

This temporal perspective alongside with the timing constraint just seen impose the need for an estimation of the complexity of the algorithm.

The first and most important aspect that drives the complexity of the algorithm is the dimension of the power set of *sentiment* and *financial variables*.

If n is the cardinality of the set of variables, its power set will contains 2^n elements. Obviously this fact together with the fully exploratory nature of our approach drives to the conclusion that the complexity of the algorithm, using the *Big O notation,* is $O(2^n)$. This fact points out that

for a full examination of the whole set of possible models one must choose carefully the set of variables that he intends to explore. Nevertheless, it is worth noting that the base set of variables that maps the *sentiment* and the financial aspects comprehends less than ten variables and the decision to include other variables would only be driven by the will of including some combination of the base set of variables. Moreover, during the testing phase it will be shown that using a proper and meaningful set of variables the temporal constraint will be satisfied.

Chapter 5

Experimental study and analysis of the results

In this chapter we will illustrate the application of the previously described methodology to the automotive industry and we will provide some proofs to the assumptions that we have considered throughout the previous chapter.

Firstly, we will cover how we have tailored the *sentiment* model to this specific industry and which financial instrument was chosen as well as the reasons behind this choice.

Secondly, we are going to present the results of an investment simulation performed by forecasting the *returns* of the chosen financial instrument exploiting the *sentiment* gathered from Twitter.

Finally, the result of the investment simulation will be compared to a benchmark in order to provide a better perspective on our results. The collection of Twitter posts started on March 2012 and the algorithm has been executed on the period that goes from the beginning of April to the end of September 2012, thus giving a fairly long period to test the properties of the proposed methodology.

5.1 Which sector?

As stated in the previous chapter, the suggested methodology is intrinsically general and can be applied to many different sectors indistinguishably. All this comes with a simple but fundamental limitation: the sector to which we are interested must be also interesting for Twitter users and thus well chatted. In fact, because of statistical significance reasons, a large enough volume of chatter is needed for each day of the year. Moreover, tweets must be quite uniformly distributed along the year in order to avoid biases in some specific periods. It follows that both topics with an average low volume of posts and those ones that peak in certain months but then disappear must be discarded. Finally, it is quite important, but not strictly required, that Twitter acts as a news medium for the chosen sector, that is, a significant number of official news (i.e. *financial* posts) should come along with *buzz* (i.e. *non-financial* posts).

Hence, among other possibilities, we opted for the automotive sector for several reasons. First, Twitter users report a sufficient number of official news, at least comparable with that of Google, and almost in real-time, as assessed during the preliminary analyses phase. Second, cars represent a topic whose influence is quite independent from the gender and not limited to a specific age range, but rather interests the whole public, from young to older people. This also implies that it would probably be well chatted, and this is the case: the levels of *buzz* related to this topic are indeed huge (cfr. Paragraph 5.6) and range from comments about new cars, commercials and key people to generic opinions about the company policy, as demonstrated by a first sample of tweets.

5.2 Which brand?

The second step is related to the choice of a proper *brand* (i.e. an index or a single stock) concerning the automotive sector. Even if the algorithm described in the previous chapter was designed to operate on a daily basis, in principle it can be also applied using a different scale (e.g. a week, a month, etc.). However, it is worth noting that not all the time scales are suitable for both types of *brand*. In particular, a single stock should not be analysed with a daily granularity, because the volume of posts of a single company, especially of those ones that convey subjectivity, will probably be too low. Therefore, a week seems to be the correct temporal horizon to be adopted. On the contrary, the total number of posts that refers to an index, that is, a composite of different companies, is equal to the aggregate of all the tweets related to each single company and thus it can be exploited for daily operations.
However, that being said, we were ultimately forced to select an index as *brand* because of a practical reason. We started the crawling phase on March 2012 with the purpose of performing an analysis on a six-month period. Hence, assuming a week as the reference time unit would have resulted in having approximately 50 points (25 weeks multiplied by 2, because a week is divided as usual into a *trading session* and the *non-trading session* of weekends) for training and testing. It stands clear that such a scenario would have been unacceptable and thus we turned to the other option, that is, an index, in order to be able to use a single day as the temporal reference.
According to what have been explained in the previous chapter, the index should fulfil some simple requirements:

- English-speaking people must be a representative sample of the set of investors of the companies that belongs to it, as only English tweets are considered for the analysis. If not, this would introduce a selection bias.

- It must comprehend a sufficient number of stocks, otherwise it would be as analysing a single company.
- A significant majority of the public-traded companies that are included in the selected index should belong to the same sector, in order to avoid tuning the model more than once.

For all the reasons exposed above, we elected *Dow Jones Automobiles & Parts Titans 30 index* (whose symbol is *DJTATO*) as the index of reference, whose stated objective is "to represent leading companies in the global Automobiles & Parts sector" and was first computed on February 12, 2001 [123]. The wording Titans 30 refers to the fact that its 30 components were selected "based on rankings by float-adjusted market capitalization, revenue and net profit" [123]. Moreover, it is a U.S. index and thus American investors are expected to be the most relevant ones. The Table 5.1 reports the complete list of *DJTATO* components and their relative weight [124].

Table 5.1: DJTATO components and their relative weight

Company	Country	Weight (%)
Daimler AG	Germany	11.41
Honda Motor Co. Ltd.	Japan	9.51
Toyota Motor Corp.	Japan	9.47
Ford Motor Co.	U.S.	6.03

Company	Country	Weight (%)
Nissan Motor Co. Ltd.	Japan	4.69
BMW AG	Germany	4.49
Johnson Controls Inc.	U.S.	4.48
Volkswagen AG	Germany	4.26
Denso Corp.	Japan	4.17
Hyundai Motor Co. Ltd.	South Korea	3.44
Bridgestone Corp.	Japan	3.43
Compagnie Generale des Etablissements Michelin	France	3.33
Suzuki Motor Corp.	Japan	2.9
Sumitomo Electric Industries Ltd.	Japan	2.77
Renault S.A.	France	2.74
Fiat S.p.A.	Italy	2.68

Company	Country	Weight (%)
Hyundai Mobis Co. Ltd.	South Korea	2.54
Porsche Automobil Holding SE	Germany	1.92
Mitsubishi Motors Corp.	Japan	1.85
Astra International	Indonesia	1.83
Toyota Industries Corp.	Japan	1.74
Genuine Parts Co.	U.S.	1.63
Harley-Davidson Inc.	U.S.	1.59
Peugeot S.A.	France	1.44
Aisin Seiki Co. Ltd.	Japan	1.34
Magna International Inc. Cl A	Canada	1.29
Yamaha Motor Co. Ltd.	Japan	0.84
Kia Motors Corp.	South Korea	0.77
Fuji Heavy Industries Ltd.	Japan	0.67

Company	Country	Weight (%)
Mazda Motor Corp.	Japan	0.64

5.3 Setting-up the crawler

Among all these companies, we decided to retain only automakers for crawling, and to exclude automotive components manufacturers. It was felt that this would not have an influence on results, since companies that produce only auto parts represent less than 25% of the index. Furthermore, if we exclude also corporations that belong to or are owned by other groups of automakers (e.g. Denso corp.) that are to be crawled, this percentage drops to less than 20%. This choice is mainly due to the fact that we did not want to overload both the crawler and the other processing components and to tune the model twice. The gain in terms of speed-up was what influenced us in solving this trade-off problem and led us to discard auto parts producers.
Hence, we decided to collect all the tweets related to the automakers included in the *DJTATO* index, following the methodology explained in the previous chapter for the choice of the most proper sets of *keywords*.

5.4 Brand model refinement

The subsequent step is the model evaluation, that is, an analysis conducted over a sample of posts in order to establish if the previously defined sets of *keywords* and *categories* are adequate in terms of precision and goodness of fit.
In particular, the first phase, also known as *noise analysis*, is aimed to evaluate the precision and recall. As explained in the previous chapter,

precision is the number of collected messages that are actually related to the *brand* over the total number of retrieved posts, while recall is the number of collected tweets that are related to the *brand* over the total number of messages that are theoretically available on the data source. It follows that a simple way to measure the precision is to manually analyse a sample of the data in order to assess if such posts are related to the *brand*. The same does not hold for recall, which is quite impossible to measure. The only viable way is to collect all the posts that are published on the data source in a given time period and to establish how many of them are actually related to the *brand* and how many of them would have been retrieved by using the chosen set of *keywords*. However, in order for this process to be effective, a big enough sample of tweets must be gathered and manually evaluated. As 175 million tweets are posted daily [8], it is evident that such an analysis is not feasible. Anyway, as a general rule of thumb, we can state that recall grows as the number of *keywords* increases. Hence, we were quite confident that recall was good enough and opted for concentrating on measuring the precision.

We collected and manually evaluated a sample of 10000 tweets posted during April 2012, thus concluding that the reached level of precision was of about 90 per cent. This value of precision was judged to be adequate for our purposes, thus no modifications were proposed.

We report here the three main errors detected during the analysis, which are all due to the weak ambiguity of the *brand name*:

- *Fiat:* in addition to the company, it can be either a Latin word (e.g. Fiat voluntas tua) or a way to refer to money that is used as the main currency of the country (e.g. fiat money, fiat currency, fiat dollar, etc.). Anyway, no alternatives were found to overcome this problem.
- *Ford:* it is an English surname. This is not a problem in general, because the number of posts that refers to common people with a

certain surname is expected to be low. However, it may be a problem if there are popular people named Ford (e.g. Harrison Ford, Robert Bruce Ford, Gerald Rudolph Ford, Francis Ford Coppola, etc.). Once again, no solutions were figured out. However, this problem seemed to be not so relevant, given the level of precision obtained.

- *Mercedes:* it can be a proper name. Also in this case, popular people with this name were found to be addressed in the sample of tweets (e.g. Mercedes Jones, a character of the TV series Glee). A possible solution was to substitute the *keyword* "Mercedes" with the complete *brand name* "Mercedes Benz" or "Mercedes-Benz". However, this choice would have led to retrieve only 168 of the 574 relevant posts related to Mercedes (i.e. only 29.27% of the total). This reduction in terms of recall would have been unacceptable and forced us to retain the same *keyword* as before.

The following step concerns the refinement of the taxonomy of *categories*. The purpose is to assess if the *categories* map in a proper way the modelled reality or if there is a lack in them. In fact, as described in the previous chapter, this taxonomy is non-domain specific and thus should be declined into a more specific one, once the sector has been chosen. In order to meet this goal, we examined 6000 further posts searching for new trending topics. This analysis led us to discover an unmapped concept that is specifically related to the automotive sector, that is, sports events (e.g. F1 and NASCAR). Hence, we decided to repeat the modelling process by introducing the *category* "Event" and by adding an ensemble of proper *semantic concepts*. On the contrary, the set of *keywords* has been kept untouched because the only way to extend it was to add the pilots' names; however, as they are quite popular, this choice would have produced the effect of dirtying the data with the inclusion of posts that are actually related to those people but not to the

brand and thus not relevant. Therefore, the *noise analysis* was not repeated and no other posts were crawled and analysed for this purpose.
This example has clearly shown that such an iterative and incremental approach has been useful to refine the model and thus it must stay at the basis of any design process.

5.5 Testing the sentiment analysis component

As we decided to rely on an external *sentiment analysis* component, we tested the quality of this tool (w.r.t. either the categorization and the *sentiment evaluation*) in terms of the three most common assessment techniques: precision, recall and accuracy.
In order to do that, we examined a sample of about 1000 posts, performing a manual *sentiment analysis*. As claimed in the Chapter 3, this task is affected by a certain level of subjectivity. Therefore, like in a previous analysis over Italian tweets, we decided to independently analyse all the messages, assigning them the most proper *category* and a value of *sentiment* and finally comparing the results until an agreement was reached. The *sentiment* was classified into positive, negative or neutral in order to be consistent with the *sentiment evaluation* component. Moreover, we retained as valid all the assumptions made during the exploratory analysis phase.
Table 5.2 reports the results obtained, where "Evaluated by the system" refers to the categorization performed by the system, while "Observed" values are those that were assigned manually. Similarly, Table 5.3 reports the results of the *sentiment evaluation*.

Table 5.2: Manual categorization results

		Evaluated by the system	
		Financial	Non-Financial
Observed	Financial	17	9
	Non-Financial	5	969

Table 5.3: Manual sentiment evaluation results

		Evaluated by the system		
		Pos	Neg	Neu
Observed	Pos	47	2	42
	Neg	0	10	8
	Neu	15	3	873

Finally, we computed the three performance measures, i.e.:

1. *Accuracy:* it is the number of items whose *category* or *sentiment* has been assigned correctly divided by all the items, that is, the sum of diagonal elements divided by the sum of all the elements. It is a quite unreliable measure, especially if categories are unbalanced [125], and this is the case.
2. *Precision:* it is the number of items whose *category* or *sentiment* has been assigned correctly and belonging to a certain class divided by all the items that are indicated by the system as belonging to that class, that is, the calculation is column-based.
3. *Recall:* it is the number of items whose *category* or *sentiment* has been assigned correctly and belonging to a certain class divided by all the available items belonging to that class, that is, the calculation is row-based.

The overall accuracy level of the categorization is 98.6%, while that of the *sentiment evaluation* is 93%. Other results are reported in Table 5.4 and 5.5.

Given the previous measurements, the *sentiment analysis* component has been judged suitable for our purposes, as it outperforms other market benchmark tools both in terms of precision and recall [117]. It is worth noting that the reached levels of performance differ from those reported in [117] either because the tool has been improved in the meanwhile and because we conducted a preliminary work aimed to find the most common errors in order to make the proper corrections.

Table 5.4: Precision and recall levels of the categorization

	Financial	Non-Financial
Precision	77.27%	99.08%
Recall	65.38%	99.49%

Table 5.5: Precision and recall levels of the sentiment evaluation

	Pos	Neg	Neu
Precision	75.81%	66.67%	94.58%
Recall	51.65%	55.56%	97.98%

5.6 Insights on returns and on Twitter posts

As we said till now, we had to face two fundamental sets of data: the *sentiment* data and the financial data. They both have been manipulated in order to obtain something useful for prediction purposes; in particular,

the posts gathered from Twitter have been analysed in order to obtain their *sentiment*, while the financial data have been collected in order to compute the daily *simple return*.
Let's focus our attention over the *sentiment* data. As a proof of what we claimed about the volumes of chatter, we report in Figure 5.1 the results of the six months period of crawling. In the whole period from the beginning of April till the end of September we have assigned a *sentiment* value to roughly 7 Millions of Twitter posts.

It is worth remembering that only English posts that have as their topic something related to the automotive industry combine to bring about the resulting volume, as stated before.

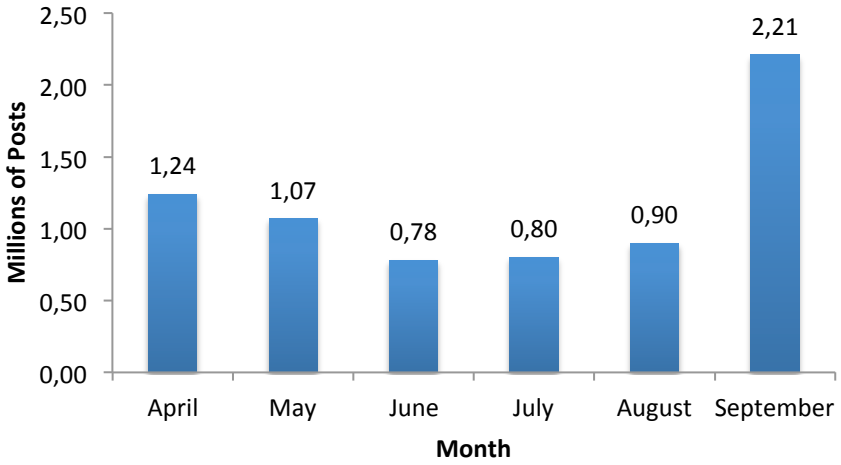

Figure 5.1: Monthly posts volume

Moreover, it stands clear that the chosen topic has been well chatted during all the months under analysis and that posts are quite uniformly distributed, with the exception of September, that presents a peak. The June-August period is characterized by a smaller amount of tweets, but this is not a problem as volumes remain sufficiently high.

Now we want to focus the attention of the reader on some descriptive statistics of the collected posts once the *sentiment analysis* had been performed. Therefore, we will provide the reader a perspective on the central tendency of the data by making use of box plots (also known as box & whiskers plots) where the four quartiles of the data are made evident together with the *outliers*, thus making graphically evident the degree of dispersion in the data. It is worth noting that we have considered *outliers* either those data points that are 1.5 times the inner quartile range *IQR* (that is the difference between the 3rd quartile and the 1st quartile) above the 3rd quartile and those data points that are 1.5 times the *IQR* lower than the 1st quartile. The legend of the graphical elements of a general box plot is given in Figure 5.2.

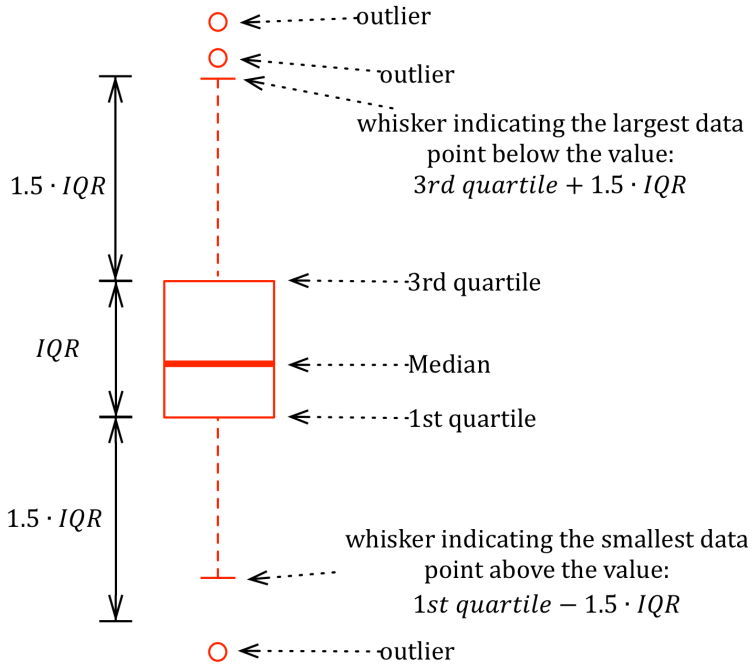

Figure 5.2: Description of the graphical elements of a box plot

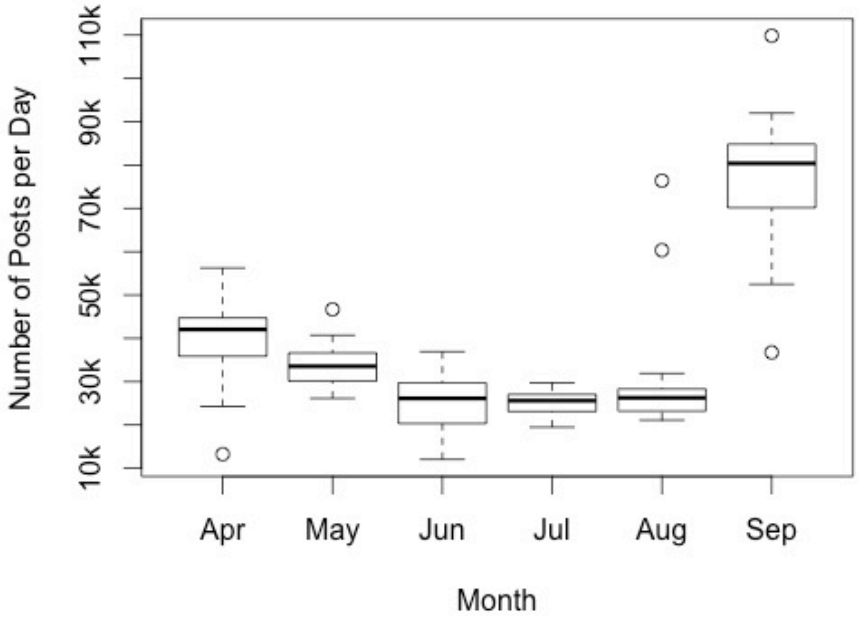

Figure 5.3: Box plot of non-financial posts

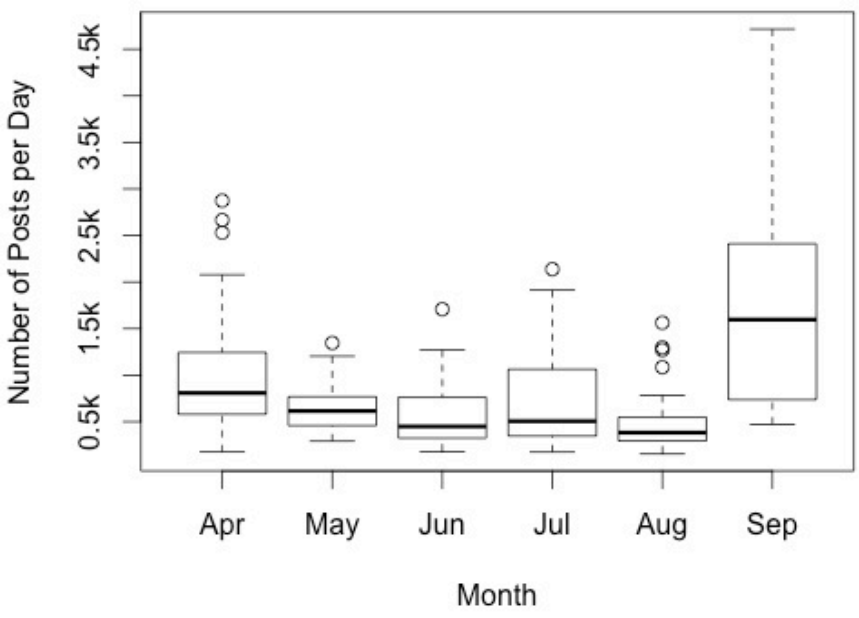

Figure 5.4: Box plot of financial posts

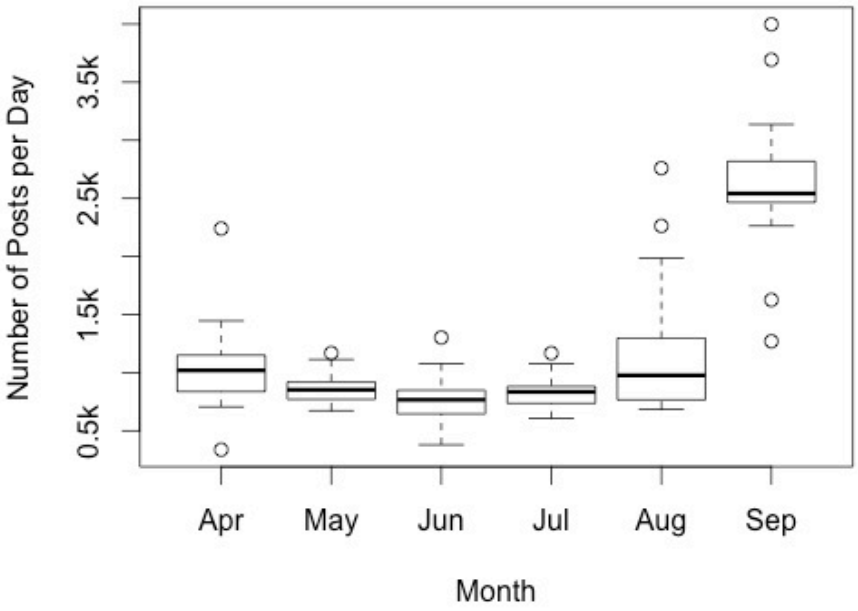

Figure 5.5: Box plot of total positive posts

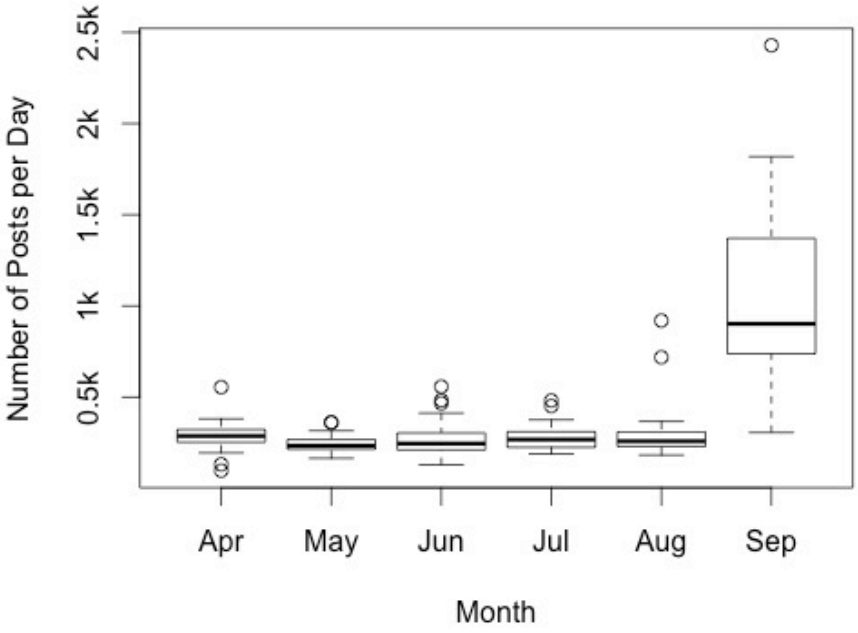

Figure 5.6: Box plot of total negative posts

It is evident that volumes of *non-financial* posts are much greater than those of *financial* posts (cfr. Figure 5.3 and Figure 5.4), as Twitter is mainly used for speaking about everyday life rather than for work or business. It follows that *buzz* cannot be ignored and should be taken into account because of the important role it plays.

Another aspect that can be easily inferred from Figures 5.5 and Figure 5.6 is that positive posts are much more than negative ones. This may be due to a general tendency of Twitter users, who prefer to share positive emotions and experiences rather than bad news or everyday problems in an attempt to project a positive image of themselves to others.

Furthermore, Figure 5.7 shows that the volumes of posts during the *night* are greater than those of the *day*, but this follows directly from the fact that the *night* period consists of more hours. However, it must be pointed out that, in spite of the limited number of hours considered, posts published during the *day* are sufficiently numerous to be considered adequate for a statistical analysis.

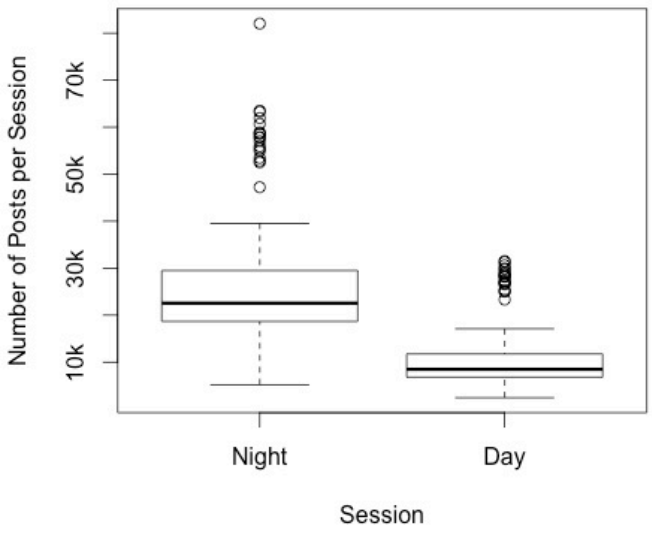

Figure 5.7: Box plot of all posts per session

Finally, the most important thing to be noticed for our purpose is that, at each level of granularity, some particular phenomena manifest in the form of *outliers*. With the expression "each level of granularity" we mean that, as shown in Figure 5.8 and Figure 5.9, data are characterized by *outliers* either if we consider only generic neutral posts or if we look specifically at positive financial posts and so on.

This is a fundamental aspect, because it is what conveys information in a predictive scenario. In fact, if the daily trend of posts had been quite uniform during the period in analysis, it would have been quite impossible to catch specific events that may have an impact on the stock market. Rather, it would have been as predicting on a random base. On the contrary, the presence of such outliers witnesses the fact that something strange happens in the data and these sudden changes should be monitored and taken into account.

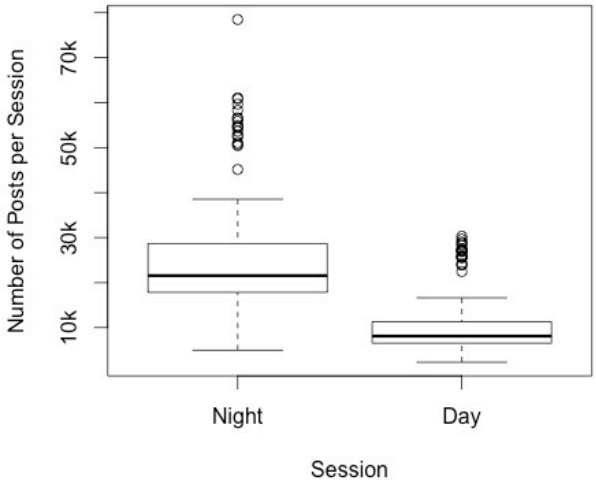

Figure 5.8: Box plot of neutral posts per session

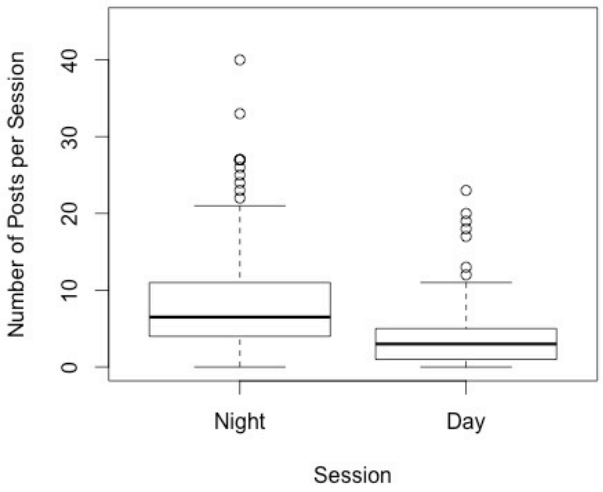

Figure 5.9: Box plot of positive financial posts per session

As a proof of what we claim, Table 5.6 reports 5 examples in which an *outlier* in terms of positive financial posts during the *night* has been immediately followed (i.e. in the subsequent *day* period) by a quite high *return*, as the average *day return* of *DJTATO* during the period under analysis has been of about 0.9%.

Table 5.6: DJTATO returns

Day	Return
April 04, 2012	-2.8%
April 19, 2012	-1.1%
April 26, 2012	1.7%
August 2, 2012	-1%
September 26, 2012	-2.5%
September 28, 2012	-1.1%

It is worth noting that the relation between peaks in chatter and in *returns* is not trivial, that is, highly positive chatter is not always

followed by a high and positive *return*, but rather it may provide some hints on the absolute value of the forthcoming *return*. However, it must be emphasized that no more than this can be inferred by a qualitative analysis of data. Therefore, the algorithm is expected to take into account this fact and to extract knowledge in order to reach the right conclusions by itself.

In addition, we point out that *outliers* really denote particular events. For instance, the increase of positive financial posts on April 04, 2012 is due to the participation of Carlos Ghosn, the Chairman and CEO of Renault and Nissan, to the New York International Auto Show, during which he acknowledged that Nissan was doing better than ever but also expressed his intention to get more U.S. share. Furthermore, he confirmed his trust in electric vehicles. Hence, users posted their comments and opinions about this fact.

The same holds for the 2 of August, during which General Motors results in terms of sales became public. In particular, people shared and commented a news according to which General Motors had topped earnings forecasts in Europe, thus generating a peak of positive *financial* posts.

Finally, Figures 5.10 and Figure 5.11 illustrate the nature of *DJTATO returns*. In particular, we can notice that *night returns* are bigger in absolute value than *day returns* and that the great majority of *returns* are concentrated around the 0, as claimed before. Table 5.7 reports also some statistical measures including *skewness* and *excess kurtosis*. We recall here that negative *skewness* indicates that the data distribution is left-skewed, that is, the left tail is longer than the right tail, while a zero *skewness* indicates a symmetric distribution (as it is the case here), whereas positive *excess kurtosis* indicates a leptokurtic distribution. A commonly accepted interpretation is that a leptokurtic distribution indicates that infrequent extreme values contribute more to the variance than frequent but modest ones, thus originating a much more acute peak

around the mean and fatter tails with respect to a Normal distribution which has an *excess kurtosis* equal to zero.

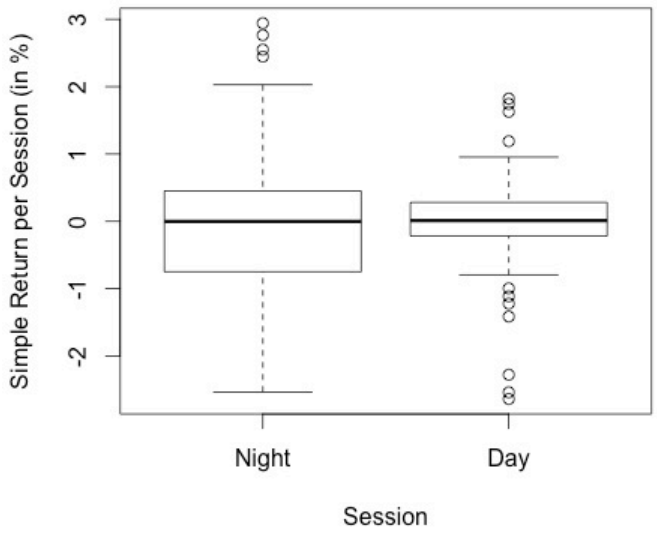

Figure 5.10: Box plot of DJTATO returns per session

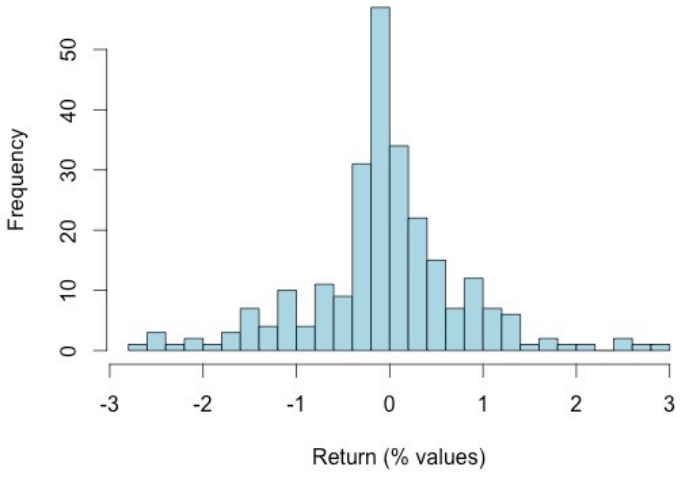

Figure 5.11: Frequencies of DJTATO returns

Table 5.7: Some statistical measures of the DJTATO returns

Mean	Variance	Standard Deviation	Skewness	Excess Kurtosis
-0.04279711	0.7092864	0.8421914	-0.04487119	1.894568

5.7 Assumptions testing on linear regression models

During the development of the algorithm we opted for dividing the whole set of identified linear regression models into some specific clusters. Firstly, we distinguished purely *financial models* from *sentiment models*, and then we further differentiated the *sentiment models* by their number of variables.

A preliminary execution of a first version of the algorithm on a subset of 80 time point yielded hundreds of different models almost without having a single repetition of them throughout the whole set of *time frames* within the sample. This fact, together with the metrics that were designed to monitor the quality of the predictions, suggested to group the models based on the criteria that we have stated above.

However, in order to validate our approach, we decided to examine the impact of each single model as well as the impact of each single class of models thus giving another perspective over the results we have obtained.

The first thing that needs to be made clear is that from here on we are going to consider the algorithm results obtained on a period consisting of 256 time points and by using all the *TFWs* ranging from 20 to 40.

A straightforward result is that the different *prediction series* originated by the use of distinct *TFWs* contain on average predictions for 87 per cent of the total number of *returns*, ranging from 81 per cent for a *TFW* of 37 points to 92 per cent for a *TFW* of 25. Hence, in this sense, their overall behaviour is quite similar, as claimed in the previous chapter. Another preliminary consideration regards the fact that the larger the *TFW* the lower the number of single unique models, as can be clearly seen in Figure 5.12.

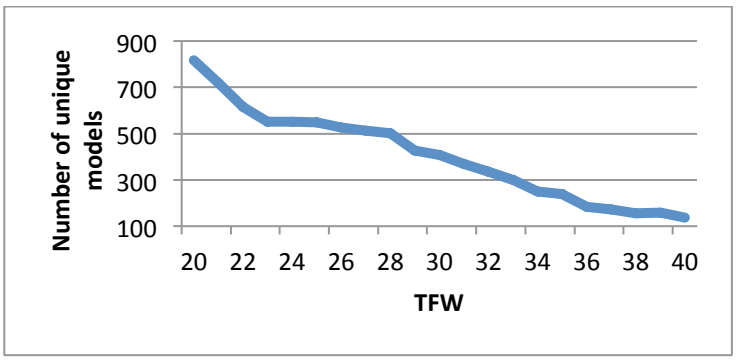

Figure 5.12: Number of single unique models per TFW

Even though the number of single unique models is great, only a small fraction of them yield more than 10% of the predictions as it is depicted in Figure 5.13. It follows that they cannot be analysed singularly. Moreover, only *POSMINUSNEG-FINANCIAL-REL* and *SENT-REL-FINANCIAL* produce more than 20% of the forecasts. These two models will be examined afterwards in this chapter, as they deserve special attention (cfr. Paragraph 5.9.2).

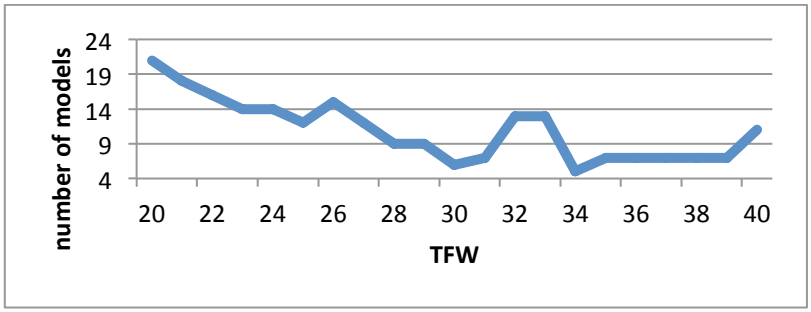

Figure 5.13: Number of single models that produce more than 10% of the predictions

If we look at the predictions made by each class of models we can see in Figure 5.14 the inversely trend of the number of predictions done by each class of models, that is, for small *TFWs* the number of predictions done by *sentiment models* is greater than that of the predictions done by *financial models*, whereas the opposite holds for bigger *TFWs*.

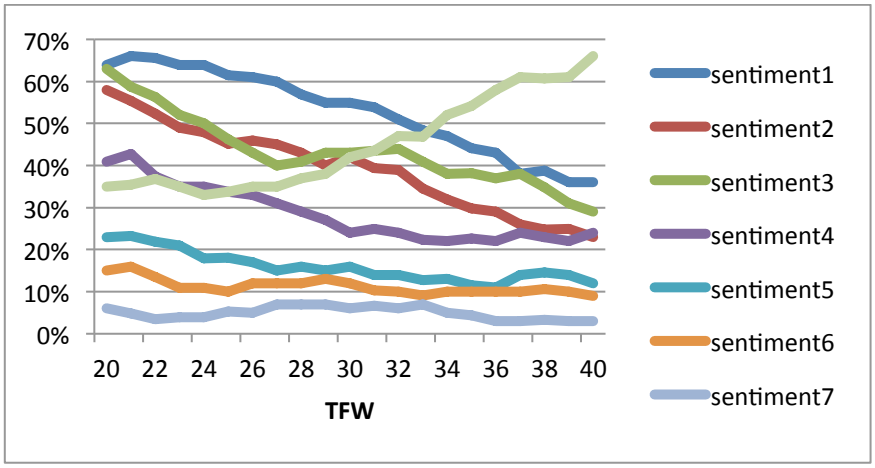

Figure 5.14: Fraction of the whole predictions yielded by each single class of model

Furthermore, we can observe that on average (the average is considered among the various *TFWs*) *sentiment models* with one, two and three variables as well as *financial models* account for more than 40% of the total number of predictions, as showed in Figure 5.15, while the others

provide less than 20% of the predictions, as anticipated in the previous chapter.

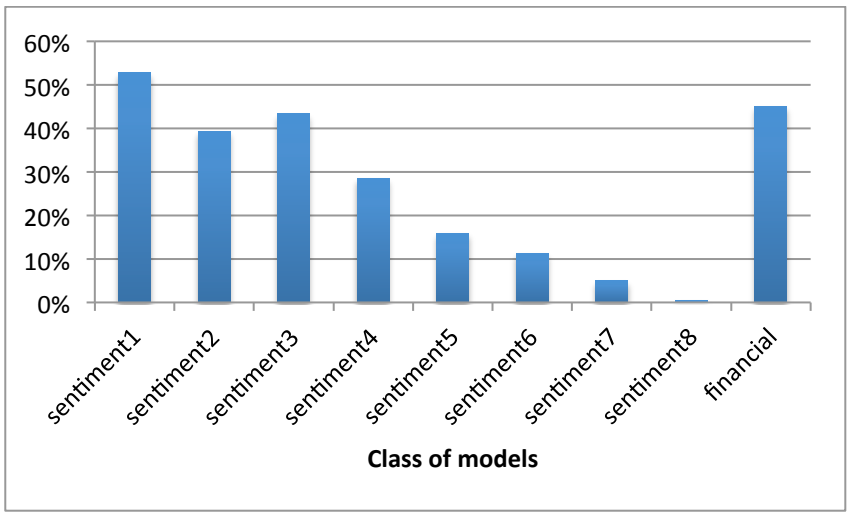

Figure 5.15: Average per cent of the total predictions per class of models

Another aspect to which is worth looking at is the one of *financial* versus *sentiment models* comparison. When both *sentiment* and *financial models* provide a prediction for a given time point, it happens that, depending on the chosen *TFW*, the sign of the predictions made by *financial models* differs from that of the predictions made by *sentiment models* in the 37 per cent of the case on average, ranging from a minimum of 25 per cent in the case of a *TFW* equal to 26 points to a maximum of 48 per cent in the case of *TFW* equal to 33. This confirms that it is worth considering the difference between *sentiment* and *financial models*.

Moreover, in order to prove our claim that no general model can be conceived, we took a practical approach. It must be pointed out here that, as hinted previously, the testing phase was commingled with the methodology development and acted as a guideline for it. In a certain sense, it can be said that our global attitude to the problem of predicting *returns* has not been top-down or model-driven, but rather bottom-up or data-driven. We started from a very basic assumption, that is, *sentiment*

expressed on Twitter has an impact on the stock market. However, we immediately questioned this and all the assumptions we made during the design process, thus taking an empirical approach. This approach (illustrated in Figure 5.16) provides that, starting from an observation of the world, each model that tries to explain that phenomenon must subsequently be tested against the reality in an iterative process that constantly involves empirical observations and the modelling phase.

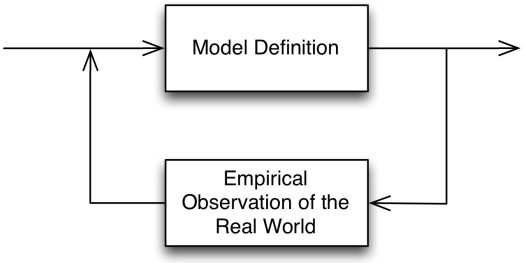

Figure 5.16: Empirical approach scheme

Therefore, both during the data collection stage and as a final proof, we tried to identify a linear regression model on the dataset gathered until that moment. In doing so, we tested all the possible combinations of the variables defined in the previous chapter. As no models on the whole dataset were found, we concluded that, as expected, the idea of a constantly changing model was the unique viable way.

5.8 Setting up the algorithm

Before running the designed algorithm, two further steps are required.
Firstly, a proper set of *variables* must be chosen in order to satisfy the temporal constraint. As explained in the previous chapter, they are divided into *financial* (i.e. the *DJTATO* 1-lagged and 2-lagged *returns*) and *sentiment variables*.

Among all the *sentiment variables* defined, we decided to retain only the most significant ones, that are:

1. POS-FINANCIAL
2. POS-NON-FINANCIAL
3. NEG-FINANCIAL
4. NEG-NON-FINANCIAL
5. NEU-FINANCIAL
6. NEU-NON-FINANCIAL
7. POSMINUSNEG-FINANCIAL-REL
8. POSMINUSNEG-NON-FINANCIAL-REL
9. SENT-REL-FINANCIAL
10. SENT-REL-NON-FINANCIAL
11. SENT-REL.

It is evident that we decided to exclude those variables that derives from the others by mean of a mere sum or difference. In fact, a sum or a difference are nothing more than particular linear regressions with coefficients equal to +/- 1 and thus the discarded variables can be theoretically obtained even if they are not pre-computed. Hence, we can conclude that this design choice does not limit or affect the analysis in any way.

The so-constructed set is made up of 13 independent variables that correspond to a total of 8192 possible combinations to be tested with a linear regression. In order to satisfy the 30 minutes time constraint we developed a parallel version of the algorithm, where a process is spawned for each single *TFW*, since the work to be done is completely independent from one *TFW* to the other. In order to give the reader an idea of the time required for the prediction of a single *return*, the multiprocessing version of the algorithm takes roughly 5 minutes of computation on a quad core processor, which fully satisfies the given time constraint.

Secondly, the three parameters α, β and γ must somehow be estimated. A possible way would be to fix them basing on a theoretical reasoning. However, as explained in the previous chapter, we preferred to keep a data-driven approach, as no reasonable theoretical assumptions could be made in this specific case. Hence, we divided the dataset into two different parts: one for the training and one for the testing. As a sufficient number of points must be provided for both tasks, but at the same time we wanted to privilege the testing phase and to avoid any overfitting phenomenon, we opted for assigning one-third of the data to the training and the rest to the testing as the best trade-off solution. Moreover, we also faced the problem of defining a proper measure that would have driven the choice of the best values for parameters. In this sense, we agreed that the most appropriate way to test our predictive algorithm was to simulate an investment scenario. Other measures such as *Mean Squared Error* (*MSE*), *Mean Absolute Error* (*MAE*), etc. seemed to be not suitable for our purposes. In fact, if the predicted *return* is very close in absolute value to the real one, but opposite in sign, the subsequent trading decision would generate a loss even if such prediction is better (w.r.t. *MSE* or *MAE*) than a prediction which has the same sign of the real *return* but a quite different absolute value.
The same holds for training, that is, we established that the best values for parameters would have been those that maximize the investment return during the training period, which were also expected to provide the largest gain on the overall period. Table 5.8 reports the results obtained. It is worth noting that history does not provide any hint in case of *financial-sentiment spread* (γ equal to 0 means that only the last value matters). The same does not hold for *EWMA* (α equal to 0.05 means that history is weighted with a 0.95 coefficient) and for β, i.e. the *QI* that drives the choice of the most proper *TFW* by successfully exploiting the history.

Table 5.8: Values obtained for parameters

Parameter Name	Optimal Value
α	0.05
β	0.4
γ	0

5.9 Simulation Results

Among the other possibilities, we assumed to invest a fixed amount of money at each time step, according to the prediction made by the algorithm. This choice is mainly due to the fact that we preferred to treat each case as independent of the others in order to be able to derive a time-independent performance measure, that is, which does not depend on the particular instant at which the investment has started.
Basing on the *return* predicted by the algorithm, three different trading decisions become possible:

- *Long:* it refers to the case in which the investor owns the security and thus gains whenever the price goes up. It follows that this option is chosen if the algorithm predicts an increase of *DJTATO*.
- *Short:* it refers to the sale of stocks that are not owned by the investor. It includes any kind of financial operation which is aimed to make profits from downturns of stocks prices. Hence, it is the option chosen if the algorithm predicts a decrease of *DJTATO*.
- *No operation:* it is the option chosen if the algorithm does not retrieve any prediction for the current *TF*.

As explained before, the investment scenario covers two-thirds of the available data, that is, the period that goes from June 12, 2012 to the end of September.

Figure 5.17 illustrates the investment results. "Our investment" indicates the cumulative performance obtained following the aforementioned approach, whereas "DJTATO" represents the benchmark, that is, the case of an investor that holds the *DJTATO* (i.e. goes long) during all the period under analysis.

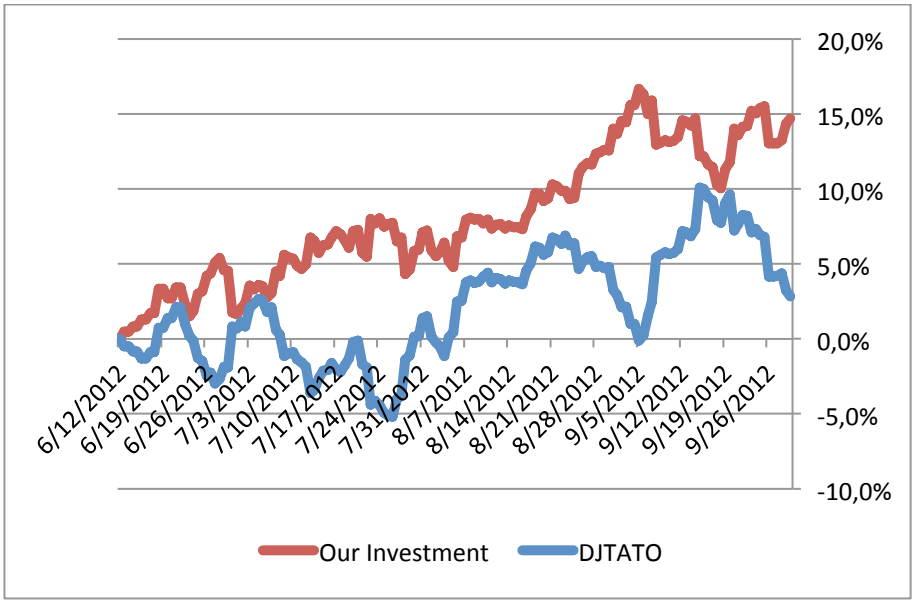

Figure 5.17: Investment results

Table 5.9 compares the performance of the two investments in terms of global *return*.

Table 5.9: Global returns of the chosen investments

Investment	Global return
DJTATO	+2.8%
Our investment	**+14.7%**

Finally, we report here three indicators that are commonly employed in order to evaluate the goodness of an investment strategy. They are:

- *Annualized return:* it is one of the best indicators, as it takes into account not only the return achieved but also the time required for getting that return. It is computed as follows:

$$AR = \left(\frac{V_f}{V_i}\right)^{\frac{YTD}{TD}} - 1$$

where V_i is the capital invested, V_f is the final capital (i.e. the initial capital plus the gain), YTD is the number of trading days in any given year (i.e. 252) and TD is the number of trading days of the investment scenario. It represents the return that would be achieved if the investment lasts a year.

- *Sharpe ratio:* it measures how well the investment rewards the investor for the risk taken. It is computed as follows [126]:

$$S = \frac{E[R_a - R_b]}{\sigma}$$

where $\mathrm{E}[\mathrm{R_a} - \mathrm{R_b}]$ is the expected value of the excess of the asset return over a benchmark return, whereas σ is the standard deviation of this expected excess return. We assumed as benchmark the 12-months *LIBOR rate* [127], which is commonly assumed as risk free rate for evaluating the Sharpe ratio.

- *Annualized volatility:* it is a statistical measure that summarizes the grade of dispersion of *returns* for a given stock or market index. A high volatility means that stock prices change rapidly, while a low volatility indicates that they are quite stable and thus it is quite impossible to obtain large returns in the short period. It is computed as follows [128]:

$$\sigma_{annual} = \sigma\sqrt{YTD}$$

where σ is the standard deviation of the logarithmic *returns*, that are defined as:

$$r_{log} = \ln\left(\frac{V_f}{V_i}\right)$$

Table 5.10 shows the results obtained. As the Sharpe ratio is far greater than 0, we can conclude that our investment strategy succeeded in outperforming the risk free investment.

Table 5.10: Some investment indicators

Annualized return	55.8%
Sharpe ratio	4.1
Volatility	13.4%

5.9.1 Statistical properties of daily P&L

The investment simulation involved 156 single *returns* whose distribution can be seen in the right box plot of Figure 5.18, whereas the spread of the profit and losses experienced by the hypothetical investor can be seen in the left box plot of Figure 5.18.

In order to better understand the shape of the distribution of either the series of the *returns* and that of the profits and losses we decided once again to point out their *skewness*.

In particular, the *skewness* of the *DJTATO returns* is 0.28, that is really close to zero but slightly positive thus indicating that the mass of the distribution is marginally concentrated on the negative side and that the right tail is slightly longer than the left one, as we can see by looking either at Figure 5.18 or at Figure 5.19. On the contrary, the *skewness* of the profit and losses (P&L) is -0.64 that is slightly negative, thus indicating that there are more positive values than negative ones and

that the left tail is longer than the right one, as it can be seen in Figure 5.18 as well as in Figure 5.19.

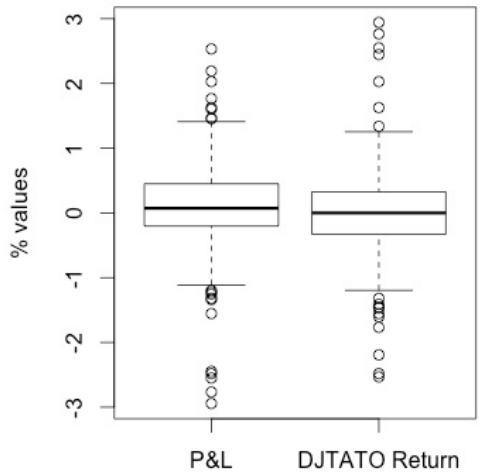

Figure 5.18: The box plots of the profits and losses of the simulation (on the left) and of the daily DJTATO returns (on the right)

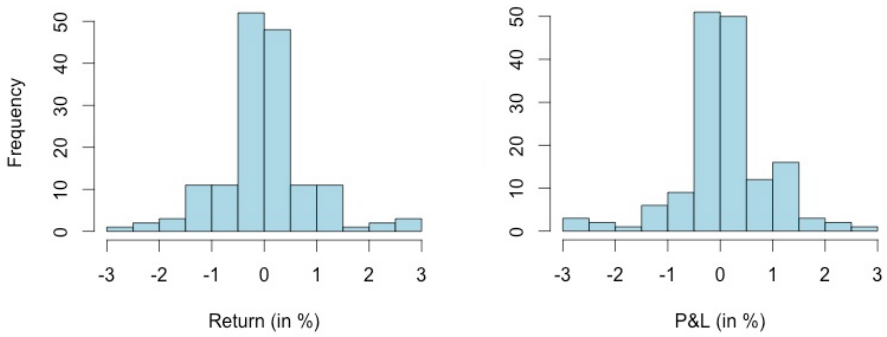

Figure 5.19: The histograms of the DJTATO returns (on the left) and of the profit and losses obtained by the simulation (on the right)

In order give the reader another perspective over the investment simulation results, it is worth noting that the number of correctly predicted *returns* is 57.5 per cent of the total number of *returns* in the time period of the simulation. However, the number of correctly

predicted *returns* is 60.9 per cent of the total number of *returns* whose absolute value is greater than 0,5% and the 65,7 per cent of the total number of *returns* whose absolute value is greater than 1%. A possible explanation of this fact may be that small *returns* are due to insignificant and unknown phenomena that cannot be captured by monitoring Twitter, while more consistent *returns* may be the consequence of particular events detectable on such social platform.

5.9.2 Where do predictions come from?

As we postponed it in Paragraph 5.7, we now want to go in depth into the analysis of the impact of the variables on the simulation of the proposed algorithm. The first thing that is clear from a superficial look at the investment simulation outcome is that *sentiment models* produce half of the predictions while the other half is due to financial models. This stated, we can evidently conclude that these two classes of models are both fundamental for predictive purposes.

As a proof of the significance of all the considered variables it is important to notice that each variable is present in at least 10 per cent of the models that originated the predictions of the simulation with the exception of *NEU-NON-FINANCIAL*. This last variable is in fact much closer to the total volume of posts during the time interval over which it is calculated without any further distinction and we can imagine that its informative content is not as interesting as that of the other variables.

As we anticipated in Paragraph 5.7, the two single variable models *POSMINUSNEG-FINANCIAL-REL* and *SENT-REL-FINANCIAL* produce more than 20 per cent of the forecasts on average (considering the whole period with 256 time points) and in fact they originate more than 24 per cent of the total number of predictions made during the simulation. This fact may point out that those posts that have been categorized as being financially relevant are indeed important for the

prediction of subsequent *returns* and thus testify the effectiveness of the categorization phase.

Chapter 6

Conclusions and future developments

With this work we decided to face an intriguing yet difficult problem, that is the prediction of future stock *returns* by using either the past series of *returns* and the *sentiment* gathered from Twitter. With respect to other similar works, some key distinguishing aspects characterize the proposed methodology.

Firstly, we have designed a *brand model* whose purpose is the mapping of all the important concepts that may be the subject of chatting on Twitter, thus driving the *sentiment analysis* in order to better understand the *sentiment* conveyed by each single post. The *brand model* has been conceived by following a top-down approach, that is we included all the aspects that we have considered to be potentially related to price swings of the *brand* related financial instruments, such as a stock or an index. However, we decided to distinguish the financially relevant features, for which we have considered different *categories*, from the *Twitter buzz* that has been conversely included in a broad *Non-Financial category*. The investment simulation on the automotive industry confirmed us that the *sentiment* conveyed on financially relevant topic is indeed important and matters as much as the *sentiment* expressed on generic issues related to the *brand*.

The second and most important distinguishing factor from other similar works is represented by the model identification approach. As there are no well-accepted methodologies for modelling the *sentiment* impact on stock markets, we did our best effort in designing an algorithm that could infer some hidden relations from the data in such a way to maximize the profits that could be made by trading the chosen financial instrument. We thus decided to identify a model over a rather limited *time frame*, that is used to predict only the subsequent *return*, and then repeating this process an arbitrary number of times by shifting forward the *time frame*; the algorithm is then in charge of selecting the most proper model by using the various quality metrics defined.

The quality of the proposed methodology has been thoroughly tested by targeting the automotive sector on a reasonable amount of time and it has been shown that the results that we have obtained outperform the benchmark, thus proving the validity of the idea of exploiting the *sentiment* as a fundamental variable.

Even though we tried our best to provide a complete analysis of the stock prediction methodology based on a properly tailored version of the *sentiment* expressed on Twitter, it will be of sure interest to delve into many aspects of the methodology. The first thing will be to apply the same methodology to different sectors such as consumer electronics, telecommunication and others, in order to further validate our approach. Yet another possible development would be the selection of different financial instruments, such as single stocks, derivatives, futures or Exchange-Traded Funds (ETFs).

Another aspect that may deserve attention could be that of augmenting the *sentiment* dimensionality. Moreover, the impact of extending the time lag of the *sentiment* variables could be explored, in such a way to infer something on the possible long lasting effect of the *sentiment*.

Additionally, a study of the nonlinearities could be managed so to discover some possible complex hidden relationships between the *sentiment* and the price deviations.

We strongly believe that the idea of exploiting the *sentiment* conveyed on Twitter for the prediction of price fluctuations of financial instruments is of sure practical interest even though it requires remarkable efforts.

Appendix

List of *semantic concepts* and *categories* to which they belong.

Category/Subcategory	Semantic concept
Hires	Hire
Hires	Hiring
Hires	Not hire
Layoffs	Fired
Layoffs	Lay-off
Layoffs	Not lay-off
Layoffs	Overstaffing
Layoffs	Sacked
Employment	Employee
Event	F1
Event	NASCAR
Factory	Facility
Factory	Production
Earnings	Earnings
Earnings	Profit
Sales	Market
Sales	Sales
Innovation	Discover
Innovation	Discovery
Innovation	Not discover
Innovation	Not patent
Innovation	Patent
Innovation	Technology

Category/Subcategory	Semantic concept
Journals	Financial magazine
Management	Management
Mergers & Acquisitions	Acquire
Mergers & Acquisitions	Acquisition
Mergers & Acquisitions	Bid for
Mergers & Acquisitions	Buyout
Mergers & Acquisitions	Merge
Mergers & Acquisitions	Merger
Mergers & Acquisitions	Not acquire
Mergers & Acquisitions	Not merge
Mergers & Acquisitions	Not subject
Mergers & Acquisitions	Takeover bid
Other deals	Agree
Other deals	Agreement
Other deals	Collaboration
Other deals	Investment
Other deals	Not agree
Other deals	Partnership
Other deals	Reorganization
Regulation	Affiliation
Regulation	Arraignment
Regulation	Basic law
Regulation	Civil law
Regulation	Consumer protection
Regulation	Court of appeal
Regulation	Decree
Regulation	Emergency law
Regulation	Export control

Category/Subcategory	Semantic concept
Regulation	Injunction
Regulation	Labour law
Regulation	Lawsuit
Regulation	Legal proceeding
Regulation	Legislation
Regulation	Licensing law
Regulation	Litigation
Regulation	Moratorium
Regulation	Ordinance
Regulation	Probation order
Regulation	Prosecution
Regulation	Ratification
Regulation	Reformation
Regulation	Regulation
Regulation	Repossession order
Regulation	Restraining order
Strikes	Not strike
Strikes	Strike

Bibliography

[1] Kaplan, A. M. and Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of social media. *Business Horizons*, 53 (1): 59-68.

[2] Benioff, M. (2012). Welcome to the social media revolution. *BBC.co.uk.* Retrieved on August, 2012 from URL: http://www.bbc.co.uk/news/business-18013662

[3] Wunsch-Vincent, S. and Vickery G. (2007). Participative web: user-created content. *Oecd.org.* Retrieved on August, 2012 from URL: http://www.oecd.org/internet/interneteconomy/38393115.pdf

[4] List of social networking websites. *Wikipedia.org.* Retrieved on August, 2012 from URL: http://en.wikipedia.org/wiki/List˙of˙social˙networking˙websites

[5] (2010). A world of connections. *Economist.com.* Retrieved on August, 2012 from URL: http://www.economist.com/node/15351002

[6] NM Incite (2011). State of the media: the social media report. *Nielsen.com.* Retrieved on August, 2012 from URL: http://cn.nielsen.com/documents/Nielsen-Social-Media-Report˙FINAL˙090911.pdf

[7] Pring, C. (2012). 100 social media statistics for 2012. *Thesocialskinny.com.* Retrieved on August, 2012 from URL: http://thesocialskinny.com/100-social-media-statistics-for-2012/

[8] Kane, B. (2012). Twitter Stats in 2012. *Webanalyticsworld.net.* Retrieved on August, 2012 from URL:

http://www.webanalyticsworld.net/2012/03/twitter-stats-in-2012-infographic.html

[9] (2012). A Case Study in Social Media Demographics. *Onlinemba.com.* Retrieved on August, 2012 from URL: http://www.onlinemba.com/blog/social-media-demographics/

[10] Chappell, B. (2012). Social Network Analysis Report - Demographic - Geographic and Search Data Revealed. *Ignitesocialmedia.com.* Retrieved on August, 2012 from URL: http://www.ignitesocialmedia.com/social-media-stats/2012-social-network-analysis-report/

[11] Schiano, D. J., Nardi, B. A., Gumbrecht, M. and Swartz, L. (2004). Blogging by the rest of us. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04: 1143-1146.

[12] Java, A., Song, X., Finin, T., and Tseng, B. (2007). Why we twitter: understanding microblogging usage and communities. *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, WebKDD/SNA-KDD '07: 56-65.

[13] Pear Analytics (2009). Twitter Study. *Scribd.com.* Retrieved on August, 2012 from URL: http://www.scribd.com/doc/18548460/Pear-Analytics-Twitter-Study-August-2009

[14] Honeycutt, C. and Herring, S. C. (2009). Beyond microblogging: Conversation and collaboration via Twitter. *Proceedings of the 42nd Hawaii International Conference on System Sciences*, HICSS '09: 1-10.

[15] Dybwad, B. (2009). Twitter Drops "What are You Doing?" Now Asks "What's Happening?". *Mashable.com.* Retrieved on August, 2012 from URL: http://mashable.com/2009/11/19/twitter-whats-happening/

[16] Odden, L. Reader Poll: How do you use Twitter? *Toprankblog.com.* Retrieved on August, 2012 from URL: http://www.toprankblog.com/2008/03/reader-poll-how-do-you-use-twitter/

[17] Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is twitter, a social network or a news media? *Proceedings of the 19th international conference on World wide web*, WWW '10: 591-600.

[18] Jansen, B. J., Zhang, M., Sobel, K., and Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, 60 (11): 2169-2188.

[19] 2010 Fortune 100 List Of Companies And Their Current Twitter Status. *Twarketing.com.* Retrieved on August, 2012 from URL: http://twarketing.com/2010/05/24/2010-fortune-100-list-of-companies-and-their-current-twitter-status/

[20] Edison Research (2011). The Social Habit II. *Edisonresearch.com.* Retrieved on August, 2012 from URL: http://www.edisonresearch.com/home/archives/2011/05/the˙social˙habit˙2011.php

[21] Owyang, J. (2008). How "Janet" Fooled the Twittersphere (and me). *Web-strategist.com.* Retrieved on August, 2012 from URL: http://www.web-strategist.com/blog/2008/08/01/how-janet-fooled-the-twittersphere-shes-the-voice-of-exxon-mobil/

[22] Mejri, O. and Plebani, P. (2012). SocialEMIS: improving emergency preparedness through collaboration. *Proceedings of the 21st international conference companion on World Wide Web*, WWW '12: 691-694.

[23] Bozzon, A., Brambilla, M., and Ceri, S. (2012). Answering search queries with CrowdSearcher. *Proceedings of the 21st international conference on World Wide Web,* WWW '12: 1009-1018.

[24] Wasserman, S. and Faust, K. (1994). *Social network analysis: methods and applications.* Cambridge University Press.

[25] Newman, M. E. and Park, J. (2003). Why social networks are different from other types of networks. *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*, 68 (3): 36122.

[26] Sysomos Inc. (2009). Inside Twitter. *Sysomos.com.* Retrieved on August, 2012 from URL:
http://www.sysomos.com/insidetwitter/

[27] Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393 (6684): 440-442.

[28] Milgram, S. (1967). The small world problem. *Psychology Today*, 2: 60-67.

[29] Leskovec, J. and Horvitz, E. (2008). Planetary-scale views on a large instant-messaging network. *Proceedings of the 17th international conference on World Wide Web*, WWW '08: 915-924.

[30] Huberman, B., Romero, D. M., and Wu, F. (2009). Social networks that matter: Twitter under the microscope. *First Monday*, 14 (1): 1-5.

[31] Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P., and Tomkins, A. (2005). Geographic routing in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102 (33): 11623-11628.

[32] Granovetter, M. (1978). Threshold models of collective behavior. *American Journal of Sociology*, 83 (6): 1420-1443.

[33] Backstrom, L., Huttenlocher, D., Kleinberg, J., and Lan, X. (2006). Group formation in large social networks: membership, growth, and evolution. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06: 44-54.

[34] Chmiel, A., Sienkiewicz, J., Thelwall, M., Paltoglou, G., Buckley, K., Kappas, A., and Hołyst, J. A. (2011). Collective emotions online and their influence on community life. PLoS ONE, 6(7): e22207.

[35] Kramer, A. D. I. (2012). The spread of emotion via Facebook. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12: 767-770.

[36] Kleinberg, J. (2008). The convergence of social and technological networks. *Communications of the ACM*, 51 (11): 66-72.

[37] Gossip protocol. *Wikipedia.org.* Retrieved on August, 2012 from URL:
http://en.wikipedia.org/wiki/Gossip˙protocol

[38] Liddy, E. D. (2001). Natural Language Processing. In *Encyclopedia of Library and Information Science*. Marcel Decker, Inc.

[39] Sentiment analysis. *Wikipedia.org.* Retrieved on August, 2012 from URL:
http://en.wikipedia.org/wiki/Sentiment˙analysis

[40] Mejova, Y. (2009). Sentiment Analysis: An Overview. *Uiowa.edu.* Retrieved on August, 2012 from URL:
http://homepage.cs.uiowa.edu/˜ymejova/publications/CompsYelenaMejova.pdf

[41] Westerski, A. Sentiment Analysis: Introduction and the State of the Art overview. *Adamwesterski.com.* Retrieved on August, 2012 from URL:

http://www.adamwesterski.com/wp-content/files/docsCursos/sentimentA˙doc˙TLAW.pdf

[42] Ozgur, A. (2002). Supervised and unsupervised machine learning techniques for text document categorization. *Master's thesis*, Department of Computer Engineering, Bogazici University, Istanbul, Turkey.

[43] Bespalov, D., Bai, B., Qi, Y., and Shokoufandeh, A. (2011). Sentiment classification based on supervised latent n-gram analysis. *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11: 375-382.

[44] McNair, D.M., Lorr, M. and Droppleman, L. F. Profile of Mood States. *Psychologyafrica.com.* Retrieved on August, 2012 from URL: http://www.psychologyafrica.com/pdf/Products/Profile%20of%20Mood%20States%20˙POMS˙.pdf

[45] Pang, B. and Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 2(1-2): 1-135.

[46] Tausczik, Y. R. and Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29 (1): 24-54.

[47] Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E., and Patwardhan, S. (2005). OpinionFinder: a system for subjectivity analysis. *Proceedings of HLT/EMNLP on Interactive Demonstrations*, HLT-Demo '05: 34-35.

[48] Brooks, I. Sentiment Discovery On Twitter. *Unt.edu.* Retrieved on August, 2012 from URL: http://students.csci.unt.edu/˜irb0005/ProjectProposalPresentation.ppt

[49] (2011). Can Twitter predict the future? *Economist.com.* Retrieved on August, 2012 from URL:
http://www.economist.com/node/18750604

[50] Balasubramanyan, R., Routledge, B. R., and Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*: 22-129.

[51] Gruhl, D., Guha, R., Kumar, R., Novak, J., and Tomkins, A. (2005). The predictive power of online chatter. *Proceedings of the 11th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '05: 78-87.

[52] Varian, H. R. and Choi, H. (2009). Predicting the present with Google Trends. *SSRN Electronic Journal.*

[53] Mishne, G. and Glance, N. (2006). Predicting movie sales from blogger sentiment. *Proceedings of AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*: 123-130.

[54] Asur, S. and Huberman, B. A. (2010). Predicting the future with social media. *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*: 492-499.

[55] Stirland, S. L. (2010). Obama's Secret Weapons: Internet, Databases and Psychology. *Wired.com.* Retrieved on August, 2012 from URL:
http://www.wired.com/threatlevel/2008/10/obamas-secret-w/

[56] Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*: 178-185.

[57] Carr, A. (2010). Facebook, Twitter Election Results Prove Remarkably Accurate. *Fastcompany.com.* Retrieved on August, 2012 from URL:
http://www.fastcompany.com/1699853/facebook-twitter-election-results-prove-remarkably-accurate

[58] Chung, J. E. and Mustafaraj, E. (2011). Can collective sentiment expressed on twitter predict political elections? *Proceedings of the 25th AAAI Conference on Artificial Intelligence*: 1770-1771.

[59] Metaxas, P. T., Mustafaraj, E., and Gayo-Avello, D. (2011). How (not) to predict elections. *Proceedings of the IEEE 3rd International Conference on Privacy, Security, Risk, and Trust* and *IEEE 3rd International Conference on Social Computing*: 165-171.

[60] Smith, A. (2011). Twitter and Social Networking in the 2010 Midterm Elections. *Pewresearch.org.* Retrieved on August, 2012 from URL:
http://pewresearch.org/pubs/1871/internet-politics-facebook-twitter-2010-midterm-elections-campaign

[61] (2011). Supercomputer predicts revolution. *Bbc.co.uk.* Retrieved on August, 2012 from URL:
http://www.bbc.co.uk/news/technology-14841018

[62] Choy, M., Cheong, M. L. F., Laik, M. N., and Shung, K. P. (2011). A sentiment analysis of Singapore presidential election 2011 using twitter data with census correction.

[63] (2011). Low expectations. *Economist.com.* Retrieved on August, 2012 from URL:
http://www.economist.com/node/18681827

[64] Equity trading. *Wikipedia.org.* Retrieved on September, 2012 from URL:

http://en.wikipedia.org/wiki/Equity˙trading

[65] Algorithmic trading. *Wikipedia.org.* Retrieved on September, 2012 from URL:
http://en.wikipedia.org/wiki/Algorithmic˙trading

[66] Aldridge, I. (2010). What is High-Frequency Trading, Afterall? *Huffingtonpost.com.* Retrieved on September, 2012 from URL:
http://www.huffingtonpost.com/irene-aldridge/what-is-high-frequency-tr˙b˙639203.html

[67] Narang, R. K. (2009). *Inside the black box: the simple truth about quantitative trading.* Wiley.

[68] Research & Programming Opportunities. *Renfund.com.* Retrieved on September, 2012 from URL:
https://www.renfund.com/vm/research˙programming.vm

[69] Teitelbaum, R. (2007). Simons at Renaissance Cracks Code, Doubling Assets. *Bloomberg.com.* Retrieved on September, 2012 from URL:
http://www.bloomberg.com/apps/news?pid=newsarchive&sid=aq33M3X795vQ&refer=home

[70] New World Order: The High Frequency Trading Community and Its Impact on Market Structure. *Aitegroup.com.* Retrieved on September, 2012 from URL:
http://www.aitegroup.com/Reports/ReportDetail.aspx?recordItemID=531

[71] Iati, R. (2009). The Real Story of Trading Software Espionage. *Advancedtrading.com.* Retrieved on September, 2012 from URL:
http://www.advancedtrading.com/algorithms/218401501

[72] Starck, P. (2008). “Black box” trading has huge potential - D.Boerse. *Reuters.com.* Retrieved on September, 2012 from URL: http://uk.reuters.com/article/2008/06/13/deutscheboerse-idUKL1366328020080613

[73] Delta neutral. *Wikipedia.org.* Retrieved on September, 2012 from URL: http://en.wikipedia.org/wiki/Delta˙neutral

[74] Mean Reversion. *Investopedia.com.* Retrieved on September, 2012 from URL: http://www.investopedia.com/terms/m/meanreversion.asp

[75] Mean reversion (finance). *Wikipedia.org.* Retrieved on September, 2012 from URL: http://en.wikipedia.org/wiki/Mean˙reversion˙(finance)

[76] Arbitrage. *Investopedia.com.* Retrieved on September, 2012 from URL: http://www.investopedia.com/terms/a/arbitrage.asp

[77] Ablan, J. (2007). Snipers, sniffers, guerillas: the algo-trading war. *Reuters.com.* Retrieved on September, 2012 from URL: http://www.reuters.com/article/2007/06/01/businesspro-usa-algorithm-strategies-dc-idUSN3040797620070601

[78] Algorithmic Trading. *Credit-suisse.com.* Retrieved on September, 2012 from URL: https://www.credit-suisse.com/investment˙banking/client˙offering/en/algorithmic˙trading.jsp

[79] Investment Banking. *Emagazine.credit-suisse.com.* Retrieved on September, 2012 from URL:

https://emagazine.credit-suisse.com/app/article/index.cfm?fuseaction=OpenArticle&aoid=363845&coid=293554&lang=EN

[80] Dark Algorithms Directory. *Advancedtrading.com.* Retrieved on September, 2012 from URL: http://www.advancedtrading.com/dir/?directory=Dark+Algorithms

[81] Heires, K. (2009). TRADING ON THE NEWS: Turning Buzz Into Numbers. *Securitiestechnologymonitor.com.* Retrieved on September, 2012 from URL: http://www.securitiestechnologymonitor.com/issues/19˙104/-23976-1.html?zkPrintable=true

[82] Van Duyn, A. (2007). City trusts computers to keep up with the news. *Ft.com.* Retrieved on September, 2012 from URL: http://www.ft.com/intl/cms/s/0/bb570626-ebb6-11db-b290-000b5df10621.html#axzz262zP23uM

[83] Nofsinger, J. R. (2005). Social mood and financial economics. *Journal of Behavioral Finance*, 6 (3): 144-160.

[84] Malkiel, B. G. (2003). The efficient market hypothesis and its critics. *The Journal of Economic Perspectives*, 17(1): 59-82.

[85] DCM CAPITAL. A Little Bit About Us... *Derwentcapitalmarkets.com.* Retrieved on September, 2012 from URL: http://www.derwentcapitalmarkets.com/about˙us/

[86] About eToro. *Etoro.com.* Retrieved on September, 2012 from URL: http://www.etoro.com/about/

[87] Tumarkin, R. and Whitelaw, R. (2001). News or noise? Internet postings and stock prices. *Financial Analysts Journal*, 57 (3): 41-51.

[88] Antweiler, W. and Frank, M. Z. (2004). Is all that talk just noise? The information content of internet stock message boards. *The Journal of Finance*, 59 (3): 1259-1294.

[89] Shtrimberg, I. (2004). Good news or bad news? Let the market decide. *AAAI Spring Symposium on Exploring Attitude and Affect in Text*: 86-88.

[90] Bollen, J., Mao, H., and Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2 (1): 1-8.

[91] Mittal, A. and Goel, A. (2011). Stock Prediction Using Twitter Sentiment Analysis. *Stanford.edu.* Retrieved on September, 2012 from URL:
http://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf

[92] Mao, H., Counts, S., and Bollen, J. (2011). Predicting financial markets: Comparing survey, news, Twitter and search engine data. *Proceedings of CoRR*.

[93] Motterlini, M. (2012). Psicofinanza: per investire spiate Twitter. *Corriere.it.* Retrieved on September, 2012 from URL:
http://www.corriere.it/economia/corriereconomia/12˙maggio˙22/motterlini-psicofinanza-investire-spiate-twitter˙3c630c2a-a407-11e1-80d8-8b8b2210c662.shtml

[94] Luss, R. and d'Aspremont, A. (2009). Predicting abnormal returns from news using text classification. *Proceedings of the 1st international workshop on advances in machine learning for computational finance*.

[95] Sehgal, V. and Song, C. (2007). SOPS: Stock prediction using web sentiment. *Proceedings of the 7th IEEE. International Conference on Data Mining Workshops*, ICDM Workshops '07: 21-26.

[96] Deng, S., Mitsubuchi, T., Shioda, K., Shimada, T., and Sakurai, A. (2011). Combining technical analysis with sentiment analysis for stock price prediction. *Proceedings of the 2011 IEEE 9th International Conference on Dependable, Autonomic and Secure Computing*, DASC '11: 800-807.

[97] Baccianella, S., Esuli, A., and Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. *Proceedings of the 7th Conference on International Language Resources and Evaluation*, LREC '10: 2200-2204.

[98] Chen, H., De, P., Hu, Y., and Hwang, B. H. (2011). Sentiment revealed in social media and its effect on the stock market. *Proceedings of the 2011 IEEE Statistical Signal Processing Workshop*, SSP '11: 25-28.

[99] Chyan, A., Hsieh, T. and Lengerich, C. (2012). A Stock-Purchasing Agent from Sentiment Analysis of Twitter. *Stanford.edu.* Retrieved on September, 2012 from URL:
http://cs229.stanford.edu/proj2011/ChyanHsiehLengerich-A˙Stock-Purchasing˙Agent˙from˙Sentiment˙Analysis˙of˙Twitter.pdf

[100] About MarketPsych. *Marketpsychdata.com.* Retrieved on September, 2012 from URL:
http://www.marketpsychdata.com/background/aboutus

[101] Applied sentiment investing using data from the social Internet. *Alphagenius.com.* Retrieved on September, 2012 from URL:
http://alphagenius.com/

[102] Pan, W., Altshuler Y. and Pentland, A. (2012). Decoding Social Influence and the Wisdom of the Crowd in Financial Trading Network. *IEEE 4th International Conference on Social Computing.*

[103] Shiller, R. J. (2003). From efficient markets theory to behavioral finance. *Journal of Economic Perspectives*, 17 (1): 83-104.

[104] Barberis, N., Shleifer, A., and Vishny, R. (1998). A model of investor sentiment. *Journal of Financial Economics*, 49 (3): 307-343.

[105] Shleifer, A. (2000). *Inefficient markets: an introduction to behavioral finance.* Oxford University Press.

[106] Krugman, P. (2009). How Did Economists Get It So Wrong? *Nytimes.com.* Retrieved on September, 2012 from URL: http://www.nytimes.com/2009/09/06/magazine/06Economic-t.html?pagewanted=1&˙r=5

[107] Barber, B. M., Odean, T., and Zhu, N. (2009). Systematic noise. *Journal of Financial Markets*, 12 (4): 547-569.

[108] De Long, J. B., Shleifer, A., Summers, L. H., and Waldmann, R. J. (1990). Noise trader risk in financial markets. *The Journal of Political Economy*, 98 (4): 703-738.

[109] Foucault, T., Sraer, D., and Thesmar, D. J. (2011). Individual investors and volatility. *The Journal of Finance*, 66 (4): 1369-1406.

[110] Yahoo! Finance. *Wikipedia.org.* Retrieved on September, 2012 from URL: http://en.wikipedia.org/wiki/Yahoo!˙Finance

[111] Mills, A., Chen, R., Lee, J. and Rao, H. R. (2009). Web 2.0 emergency applications: how useful can Twitter be for emergency response. *Journal of Information Privacy & Security*, 5(3):3–26.

[112] Nonlinear system. *Wikipedia.org.* Retrieved on September, 2012 from URL: http://en.wikipedia.org/wiki/Nonlinear˙system

[113] Linear system. *Wikipedia.org.* Retrieved on September, 2012 from URL: http://en.wikipedia.org/wiki/Linear˙system

[114] Wilson, T., Wiebe, J., and Hoffmann, P. (2005b). Recognizing contextual polarity in phrase-level sentiment analysis. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05: 347-354.

[115] Purcher, J. (2010). iPhone + Logo - Now a Registered Trademark of Apple Inc. *Patentlyapple.com.* Retrieved on September, 2012 from URL:
http://www.patentlyapple.com/patently-apple/2010/02/iphone-is-now-a-registered-trademark-of-apple-inc.html

[116] Precision and recall. *Wikipedia.org.* Retrieved on September, 2012 from URL:
http://en.wikipedia.org/wiki/Recall˙(information˙retrieval)

[117] Barbagallo, D. (2010). A data quality based methodology to improve sentiment analyses. *Research doctoral program*, Department of Electronics and Information, Politecnico di Milano, Milan, Italy.

[118] NIST/SEMATECH (2012). e-Handbook of Statistical Methods. *Itl.nist.gov.* Retrieved on September, 2012 from URL:
http://www.itl.nist.gov/div898/handbook/

[119] Harrell, F. E. (2001). *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis.* Springer.

[120] Tsay, R. S. (2005). *Analysis of financial time series.* Wiley.

[121] Moving average. *Wikipedia.org.* Retrieved on September, 2012 from URL:
http://en.wikipedia.org/wiki/Moving˙average

[122] Stiver, J. D. Some Techniques Used in Technical Analysis. *Nd.edu.* Retrieved on September, 2012 from URL:

http://www.nd.edu/~jstiver/FIN475/tech.pdf

[123] Dow Jones Automobiles & Parts Titans 30™ Index. *Djindexes.com.* Retrieved on May, 2012 from URL:
http://www.djindexes.com/mdsidx/downloads/fact˙info/Dow˙Jones˙Automobiles˙and˙Parts˙Titans˙30˙Index˙Fact˙Sheet.pdf

[124] Dow Jones Automobiles & Parts Titans 30 Index (DJTATO). *Wikinvest.com.* Retrieved on October, 2012 from URL:
http://www.wikinvest.com/index/Dow˙Jones˙Automobiles˙%26˙Parts˙Titans˙30˙Index˙(DJTATO)

[125] Potts, C. (2011). Sentiment Symposium Tutorial: Classifiers. *Christopherpotts.net.* Retrieved on November, 2012 from URL:
http://sentiment.christopherpotts.net/classifiers.html

[126] Sharpe ratio. *Wikipedia.org.* Retrieved on November, 2012 from URL:
http://en.wikipedia.org/wiki/Sharpe˙ratio

[127] LIBOR - current LIBOR interest rates. *Global-rates.com.* Retrieved on November, 2012 from URL:
http://www.global-rates.com/interest-rates/libor/libor.aspx

[128] Cabrera, L. F. Calculating Historical Volatility (HV) with Example. *Lfrankcabrera.com.* Retrieved on November, 2012 from URL:
http://www.lfrankcabrera.com/calc-hist-vol.pdf

[129] Fondi di investimento. *Ilsole24ore.com.* Retrieved on October, 2012 from URL:
http://www.ilsole24ore.com/finanza-e-mercati/fondi-24.shtml?refresh˙ce