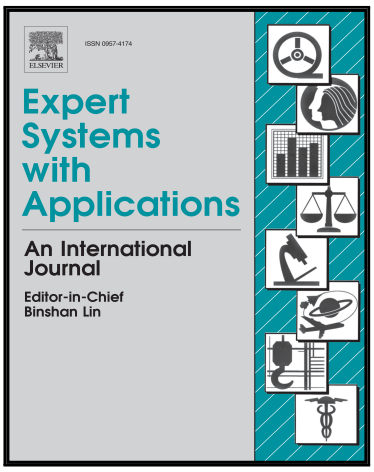

Company Event Popularity for Financial Markets Using Twitter and Sentiment Analysis

Mariana Daniel , Rui Ferreira Neves , Nuno Horta

PII: S0957-4174(16)30657-1
DOI: 10.1016/j.eswa.2016.11.022
Reference: ESWA 10993

To appear in: *Expert Systems With Applications*

Received date: 19 April 2016
Revised date: 3 November 2016
Accepted date: 18 November 2016

Please cite this article as: Mariana Daniel , Rui Ferreira Neves , Nuno Horta , Company Event Popularity for Financial Markets Using Twitter and Sentiment Analysis, *Expert Systems With Applications* (2016), doi: 10.1016/j.eswa.2016.11.022

Highlights

- The work proposes an Event Popularity Algorithm for Financial Trading.
- The approach is based on sentiment analysis to the social network Twitter.
- Planning and performing a financial community for the extraction of analyzed tweets.
- The events are focused on the thirty companies that compose the Dow Jones index.

COMPANY EVENT POPULARITY FOR FINANCIAL MARKETS USING TWITTER AND SENTIMENT ANALYSIS

Mariana Daniel[a] (marianadaniel7@gmail.com)
Rui Ferreira Neves[a] (rui.neves@tecnico.ulisboa.pt)
Nuno Horta[a] (nuno.horta@lx.it.pt)

[a] Instituto de Telecomunicações
Instituto Superior Técnico
Lisboa, Portugal

ABSTRACT

The growing number of Twitter users makes it a valuable source of information to study what is happening right now. Users often use Twitter to report real-life events. Here we are only interested in following the financial community. This paper focuses on detecting events popularity through sentiment analysis of tweets published by the financial community on the Twitter universe. The detection of events popularity on Twitter makes this a non-trivial task due to noisy content that often are the tweets. This work aims to filter out all the noisy tweets in order to analyze only the tweets that influence the financial market, more specifically the thirty companies that compose the Dow Jones Average. To perform these tasks, in this paper it is proposed a methodology that starts from the financial community of Twitter and then filters the collected tweets, makes the sentiment analysis of the tweets and finally detects the important events in the life of companies.

1. INTRODUCTION

Due to the speed of change and the increase of complexity in the financial markets it is necessary to create technological tools that help investors to correctly apply their assets in order to achieve a significant profit. Since the stock markets represent to investors, a form of profitability, in fact this type of investment has a high associated risk making it impossible to guarantee a certain profit. Price forecast as well as the right time to invest or to exit the stock market is a much discussed topic by investors.

Several techniques are known and applied by investors in order to increase profit and minimize risk. One of the most used is Technical Analysis that is based on the hypothesis that past patterns that relate to behavior of prices in each asset tends to repeat itself in the future (Gorgulho, Neves, & Horta, 2011). On the other hand, Fundamental Analysis is another type of analysis, also used by investors, that looks at macro indicators or the balance sheet of a company (Deschatre, 2009). It is also through the results of the Fundamental Analysis that can be determined whether an asset is overvalued or undervalued.

The greater or lesser supply/demand of a certain stock can be influenced by the historical behavior of prices, future prospects related to the performance of the company issuing of stock or even by news published in blogs, social networks.Studies by (Hirshleifer, 2001), (Chan, 2003), (Vega, 2006) and (Tetlock, 2007) have shown that both the informational aspects as the affective aspects of the text of a news may affect the financial markets both in terms of the impact on business volumes and on stock prices as the level of volatility. The analysis of affective aspects on particular news is done through a technique called Sentiment Analysis, which is crucial to understand if news is positive or negative and can in some way influence financial markets.

Today, the Internet is used to find information but also to share information, share knowledge and also serve as a channel for business. Online social networks are part of millions of people worldwide every day, becoming over the years an important communication platform that brings together various information, including opinions and feelings expressed by users in sharing content in news messages published. Social networks have attracted the attention of many researchers who aim to correlate the content of the various publications with the events of real life (Sayyadi, Hurst, & Maykov, 2009). The large interest happens because many events are published by the traditional media with a delay, while on social networks the event is published almost as instantly. Thus, the big question is to what extent the information provided on social networks truthfully reflects the actual events and is it possible to use this information to detect events.

The objective of this work is to develop a system for the detection and discovery of the popularity of special events in the financial area from social networks, more specifically the social network Twitter. As this proposed work is focused on the stock market we are only interested in detecting major events that can change the evolution of that specific stock and we only care if the news is positive or negative. Twitter is a very noisy source of news. The proposed work provides a basis for event popularity that have to deal with massive amounts of data and very high levels of noise typical of social networks. Twitter is an online social network that allows users to send and receive personal updates from other contacts (texts up to 140 characters, known as "tweets"). The system for event popularity focuses on the events in the life of the companies that make up the Dow Jones index. The methodology consists of several stages, from the construction of the financial community Twitter, through the collection of tweets, the analysis of the

content of tweets until in a final step the events are detected and analyzed. These real events that we are looking for are events that must demonstrate like or dislike through the comments posted about a particular company by the defined financial community. Examples of such events: product launches, special events in the life of companies, campaigns, among others.

This paper presents an original work with respect to event popularity. Here we study the events that are directly related to the thirty companies that comprise the Dow Jones index and the opinion of the financial users of each company. There are four main contributions. First the algorithm that defines a financial community inside Twitter and extracts all tweets for each user. Then the tweets filtering algorithms that allow us to get only the tweets related to the financial market more properly the thirty companies that compose the Dow Jones index. The third contribution of this paper is the creation of a simple sentiment analysis tools and the conjunction with other three known sentiment analysis algorithms. Through the four algorithms we evaluate the tweets individually in positive, negative and neutral scores. For every company the algorithms create a daily score that relates to the sum of the scores of all the tweets that occurred that day. The last contribution of this paper is the manual detection of events that have been characterized by large positive peaks and large negative peaks and are directly related to special events that positively and negatively affected the financial community.

The presented paper provides a detailed discussion on a new approach for event popularity through the social network Twitter. The paper is structured as follows: Section 2 addresses the theory behind the developed work, namely some concepts relating to market analysis as technical analysis and fundamental analysis. Also in this section are some tools that are part of sentiment analysis, a much studied method nowadays to be able to extract the sentiment of news and comments published on social networks. Finally in this section presents some literature concerning the social network Twitter, their influence on the market and still some works produced in the area of event detection in social networks focusing on Twitter. Section 3 illustrates the system architecture and respective function of each module. Section 4 proposes the validation procedure used to evaluate the developed strategy, where three case studies are presented that prove the event popularity by the proposed methodology. Section 5 summarizes the provided document, supplies the respective conclusion and the proposal of possible future work.

2. Literature Review

2.1 Market Analysis

The concept of efficient markets was proposed by (Fama, 1970), and subsequently conducted several studies in an attempt to test the theory. This theory is one of the most important and controversial issues within the financial theory, on the one hand, the proponents of this hypothesis and on the other, critics and opponents. According to this hypothesis, the market is considered efficient when the prices of financial products quickly reflect any change in the information available on the market, preventing the achievement of unusual earnings. The Random Walk Theory, (Singal, 2006) , argues that you cannot look to the past movements of a stock, pattern or trend to predict future market moves. The market acts irrationally, and the unpredictable movements of prices, following a "random walk" as well defined Maurice Kendall, the creator of this theory. According (Haugen, 2001) the behavior of prices in efficient

markets is the product of rational behavior of investors, while in inefficient markets, the price behavior results from the emotional and psychological state of stakeholders.

Meanwhile some investors believe that you can beat the market using Fundamental Analysis or Technical Analysis (Silva, Neves, & Horta, 2015). Fundamental Analysis evaluates a business from its economic and financial information. Knowing the value of a company from its numbers, you can compare this value with the value that the market assigns to it. These numbers are derived from the overall economy, the particular industry's sector, or most typically, from the company itself. Figures such as inflation, joblessness, Return on Equity (ROE), Debt levels, and individual Price to Earnings (PE) ratios can all play a part in determining the price of a stock (Cunningham, 1997).

Another approach to analyze the stock market and the future evolution of stock prices is the Technical Analysis. In 90 years, this market analysis method was introduced by Charles Henry Dow, one of the creators of the famous Dow Jones Industrial Average (DJIA) and founder of "The Wall Street Journal." To build his theory (Brown, Stephen, Goetzmann, & Kumar, 1998), Dow, was based on three main assumptions:

- ✓ The price is a comprehensive reflection of all market forces. At any given time, all market information and their strengths are reflected in prices;
- ✓ Prices move in trends that can be identified and turned into profit opportunities;
- ✓ The price movements are historically repetitive.

As mentioned above, many investors have been investigating a way to predict future prices using a variety of algorithms that use fundamental analysis and or technical analysis. These tools are used by professionals or amateur speculators to analyze the movement of prices of some financial assets. The main factor that influences the stock price is demand and supply. But there are other aspects that influence the stock price. In the information age, news can spread around the world, sometimes at a higher speed than it happens. However, the news is not the only means of dissemination of information that influence the financial market. Reviews and publications that daily invade social networks also allow us to extract very specific information about the type of motivations of the people who produce them. These motivations may have implicit positive or negative feelings. Today new indicators can be created based on the information found on the Internet more specifically on social networks. It is the objective of this paper to use the information from social networks to create new indicators that could be used to detect corporate events in the stock market.

2.2 Sentiment Analysis on Text Classification

The wide expansion of the Internet generates different information about several subjects and frequently this information implicitly contains opinions. (Indurkhya & Damerau, 2010) mention that the opinions are so important that, wherever they want to make decisions, people want to hear the opinion of others. This is not only true for people, but also for the organizations that have seen the notion of customer opinion about their products and services as an added value to organizations.

The analysis of sentiments or opinion mining is the computational study of opinions, feelings and emotions expressed in text. This technique has been widely studied as a tool in the repertoire of social media analysis carried out by companies, marketers, and political analysts, (Taboada, 2016). (Milagros,

Tamara, Jonathan, Enrique, & Francisco, 2016) proposed a new approach to predict sentiment in informal texts using unsupervised dependency parsing. The authors have implemented an algorithm based on sentiment propagation using linguistic content without training. The results confirmed the competitive performance and the robustness of the system.

In order to explore the area of Opinion Mining, (Balazs & Velásquez, 2016), presented a survey on Information Fusion applied to this area. The authors presented a survey of the most popular Opinion Mining techniques, defined the Information Fusion field. Information Fusion is the field charged with researching efficient methods for transforming information from different sources into a single coherent representation. This paper proposed a framework for guiding the fusion process in an Opinion Mining system and reviewed some of the studies that have successfully implemented Information Fusion techniques in the Opinion Mining context.

The use of data mining techniques to predict the financial markets has been extensively studied in numerous publications. (Schumaker & Chen, 2009) presented a study with the objective of finding the actual price of stocks listed on the S&P 500 using a SVR algorithm by applying text mining techniques in financial news articles. The proliferation of online documents and texts published by users led to a recent increase in the area of sentiment analysis and its relationship with the financial markets. A recent paper (Geva & Zahavi, 2014) published in order to evaluate the effectiveness of augmenting numerical market data with textual-news data, using data mining methods, for forecasting stock returns in intraday trading. Another paper that proves the impact of financial news articles on stock price return it was published by (Xiaodong, Haoran, Chen, Jianping, & Xiaotie, 2014) . In this paper, the authors analyzed the news impact from sentiment dimensions. At the first stage the authors implemented a generic stock price prediction framework and in the second stage they used two different dictionaries to construct the sentiment-dimensions. The authors evaluated the models prediction accuracy and empirically compare their performance at different market levels. The results shown that the sentiment analysis does help improve the prediction accuracy. But, simply focusing on positive and negative dimensions could not bring useful predictions. Finally the authors concluded that there is a minor difference between the models using different sentiment dictionaries.

Several studies have presented an overview of some techniques used in sentiment classification. Lexical resources for sentiment analysis have attracted a great interest from the computational linguistics community. (Bradley & Lang, 1999) released ANEW, a lexicon with affective norms for English words. The application of ANEW to Twitter was explored by (Nielsen, 2011), leveraging the AFINN lexicon. (Jain & Nemade, 2010) labeled a list of English words in positive and negative categories, releasing the Opinion Finder lexicon. The development of lexicon resources for strength estimation was addressed by (Thelwall, Buckley, & Paltoglou, 2012), leveraging SentiStrength. (Esuli & Sebastiani, 2006) and later (Baccianella, Esuli, & Sebastiani, 2010) extended the well known Wordnet lexical database (Miller, Beckwith, Fellbaum, & Gross, 1990) by introducing sentiment ratings to a number of synsets, creating SentiWordnet.

2.3 TWITTER

Because of the growing of social networks over the years, some studies focus on the application of sentiment analysis tools in publications, comments and articles posted by users. Social networks have been standing out in the grand universe that is the Internet due to accelerate communication, as they allow anyone to become a content producer. A research (Chen, De, Hu, Jeffrey, & Hwang, 2013) indicates that most financial market professionals and clients use social networks for professional reasons. This is a revealing trend that social networks have a significant weight in influencing the markets. Today, Twitter is a great source of information for this type of work. So Twitter has been winning more and more space allowing a faithful behavioral picture of the individual and his relationship group which makes it a great source for analysts. (Bollen, Maoa, & Zengb, 2011), investigated whether measurements of collective mood states derived from large-scale Twitter feeds are correlated to the value of the Dow Jones Industrial Average (DJIA). They analyze the text content of daily Twitter feeds by two mood tracking tools, namely OpinionFinder that measures positive vs. negative mood and Google-Profile of Mood States (GPOMS) that measures mood in terms of 6 dimensions (Calm, Alert, Sure, Vital, Kind, and Happy).

The intentions of Twitter users are different. Some people use Twitter only as a means of conversation, to talk about their daily activities, others use it for professional reasons, finally others use it to disseminate malicious content. To understand the influence of users on Twitter, (Yang, Steve Y.; Mo, Sheung Yin Kevin; Zhu, Xiaodi, 2014) published a study that consists of forming a financial community on Twitter where users of this community share interests in the financial market. The results illustrate that a strong interdependence between the social climate and the movement of stock prices by creating the financial community can be created. This study could be concluded that the sentiment generated by each node of the financial community has predictive power in consistent market returns and market volatility.

With the restriction to 140 characters that Twitter imposes to post a tweet, sometimes many users can't express themselves in the best way so they use emoticons. So with fewer letters, users can express feelings such as happiness, disgust, anger, shyness, and others. In the following article, the authors use specific emoticons to form the training set for sentiment classification (Go, Bhayani, & Huang, 2009). They present a new approach to automatically classify the sentiment of Twitter messages. These messages are classified as either positive or negative with respect to a query term. The training data consists in Twitter messages with emoticons which are used as noisy labels. The approach of this work is to use different machine learning classifiers and extractors of resources.

There are wide spread discussions and research about web forum, blogs and twitters as alternative form of political debate. Some researchers have acknowledged the quality of the more prominent political blogs while others doubted the capabilities of the blogs to aggregate and convey the information. Several case studies have found that the online information has been quite successful acting as indicator for electoral success. A paper published by (Wang, Can, Kazemzadeh, Bar, & Narayanan, 2012) describes a system for real-time analysis of public sentiment toward presidential candidates in the 2012 U.S. election as expressed on Twitter. With this analysis, they seek to explore whether Twitter provides insights into the unfolding of the campaigns and indications of shifts in public opinion. The design of the sentiment model used in their system was based on the assumption that the opinions expressed would be highly subjective and contextualized.

Today new indicators can be created based on information found on the Internet, specifically on social networks. The information extracted from social networks to create new indicators can be the key to detect important events both in people's lives and in the lives of various organizations. By detecting corporate events in organizations directly may be able to detect future fluctuations in stock market depending on the type of event and the kind of feeling that is implicit in the event. In the next section we present some work in the area of event detection through the social network Twitter.

2.4 EVENT DETECTION

The social networking activity rates have increased a lot over the years. Hundreds of millions of users are registered in these social networks such as Twitter. Users exchange thoughts and count episodes of day-to-day across tweets that share. Often describe events in real-time in various parts of the world. Tweets increasingly have proven very useful in the event detection and feeling associated with events that have occurred. Hence the sentiment analysis is an indispensable tool when the goal is to detect events on social networks and discover if the user has positive or negative information to transmit.

The detection of natural disasters and social events using the social network Twitter has been widely analyzed and discussed by researchers. These events often have several properties: i) are large-scale, with many users interested in experiencing the event and ii) influence the daily lives of people for various reasons and for this reason they are induced to post a tweet about the event. Such events include social events, such as large parties, sporting events, exhibitions, promotion of products, accidents and political campaigns. They also include natural events such as storms, heavy rain, tornadoes, typhoons / hurricanes / cyclones and earthquakes. (Winerman, 2009) states that as an event occurs that causes panic, people seek information on social networks. This article cites the example of the tragedy of Virginia Tech where students were able to formulate a complete list of all the deceased students one day before the authorities.

As we already know, social networks have become the main sources of information on real-world events. Most approaches that aim at extracting event information from such sources typically use the temporal context of messages. However, exploiting the location information of georeferenced messages, too, is important to detect localized events, such as public events or emergency situations. Users posting messages that are close to the location of an event serve as human sensors to describe an event. With the facility that Twitter has to filter the location of publication, many studies use Twitter as a data source in order to obtain a large number of data by location and thus be able to detect events in a particular city or country.

Several articles have been published in the context of exploring and researching techniques for event detection from Twitter streams. These techniques are aimed at finding real events that unfold in space and time. It has been a great challenge get techniques to detect events on Twitter able to handle large amounts of information without meaning and full of noise content. (Atefeh & Khreich, 2015) presented a new approach to classify techniques according to the event type, detection task and method, and discusses commonly used features.

(Gomide, Lima, Gomide, Roque, & Silva, O Twitter como Instrumento de Detecção de Epidemias de Dengue e Desenvolvimento de, 2014) published a study in order to examine the usefulness of Twitter as possible dengue outbreaks detection tool in the development of public policies in Brazil. With this study it

was shown that Twitter has demonstrated potential as dengue epidemics detection tool for the analysis showed that the Twitter data have the same behavior as compared to the data provided by the Ministry of Health. The behavior of people on social networks during events or emergencies has also been the subject of research. (Mendoza, Poblete, & Castillo, 2010) and (Starbird & Palen, 2010), the authors determined how information was disseminated throughout the network via retweets of news for two natural disasters, the flooding of the Red River and fires in Oklahoma. Messages posted on Twitter have also been used to predict the occurrence of earthquakes in (Sakaki, Okazaki, & Matsuo, 2010) and (Lampos & Cristianini, 2012). Sakaki et al. developed techniques for identifying earthquake events on Twitter by monitoring keyword triggers (e.g., "earthquake" or "shaking"). In their setting, the event must be known a priori, and should be easily represented using simple keyword queries. (Sankaranarayanan, Samet, Teitler, Leiberman, & Sperling, 2009) identified late breaking news events on Twitter using clustering, along with a text-based classifier and a set of news "seeders," which are handpicked users known for publishing news (e.g., news agency feeds). (Petrović, Osborne, & Lavrenko, 2010) used locality-sensitive hashing to detect the first tweet associated with an event in a stream of Twitter messages.

A recent paper in the field of sentiment analysis was published by (Fernández-Gavilanes, Álvarez-López, Juncal-Martínez, Costa-Montenegro, & González-Castaño, 2016) which basically created an approach to detect the sentiment of short messages such as tweets, SMS and comments. The approach that created those features is unsupervised, and can be applied to different domains. The authors based on the determination of dependencies between lemmatized tagged words using a sentiment propagation algorithm that took into account and distinguished between key linguistic phenomena, namely, intensification, modification, negation and adversative and concessive relations. This study has shown more sophistication and precision compared to simplistic approaches which assign a polarity to a word as has been done in most studies.

Compared to various studies that have been done based on generic dictionaries with numerical polarities assigned to words that are found on the Internet, this approach provides better accuracy compared with these studies. This accuracy comes from the fact that this approach does not use dictionaries already defined and instead applies a context-based algorithm to automatically create the dictionaries each particular context. The experiences of this study showed that the newly created lexicons are superior in sentiment prediction using unsupervised approach.

As the literature presented above has given to realize that the Twitter messages reflect useful event information for a variety of events of different types and scale. These event messages can provide a set of unique perspectives, regardless of the event type, reflecting the points of view of users who are interested or even participate in an event. In particular, for unplanned events, Twitter users sometimes spread news prior to the traditional news media. Even for planned events, Twitter users often post messages in anticipation of the event, which can lead to early identification of interest in these events. Additionally, Twitter users often post information on local, community-specific events, where traditional news coverage is low or nonexistent.

The following table (Table 1) shows the revision of some important literature in this section, referencing to each work, the simulation period, the main algorithms used and still the best results obtained.

Table 1 – Table of recent pattern recognition papers and their performance.

Reference	Date	Period of Simulation	Data Collected	Market Tested	Algorithms	Opinion Mining Tools	Best Results
Bollen et al.	2011	28 Feb 2008 To 19 Dec 2008	Tweets	DJIA	SOFNN	Opinion Finder & GPOMS	Accuracy (%) = 86.7
Yang, et al.	2013	15 Oct 2013 To 15 Nov 2013	Tweets	DIJA	Algorithm developed	Dictionary of (Hill & Liu, 2004)	Sentiment by critical nodes in the financial community
Go et al.	2009	6 April 2009 To 25 June 2009	Tweets	-	MaxEnt	Unigram + Bigram	Accuracy (%) = 83
Wang, et al.	2012	12 Oct 2011 To 29 Feb 2012	Tweets	DIJA	Naïve Bayes	Multi-dimensional Model	Accuracy (%) = 59
Gomide et al.	2014	1 Dec 2010 To 31 May 2011	Tweets	Brazil	Alpha de Cronbach / Spearman Correlation and Cluster Analysis	Algorithm developed by authors	Spearman correlation = 0.924 (high positive correlation)
Mendoza et al.	2010	27 Feb 2010 To 2 March 2010	Tweets	Chile	Algorithm developed with Hashtag's	-	Veracity of tweets (%) = 95.5

3. System Architecture

The objective of the proposed system is to detect and find the popularity of special events on Twitter that may influence the financial markets. The main indicator for the implementation of the model consists of the extraction of the feeling implicit in tweets posted by users on the social network. This indicator is based in four text analysis tools that return daily sentiment expressed in Tweets. This feeling oscillates between positive, neutral and negative relative to a score calculated depending on the mood of the tool used. Basically a high score indicates the presence of feelings that express positivity about the company. It is important to understand why such a high score, so that is necessary to understand the event that occurred on this day that raised positive feedback between users. By contrast one day composed of extremely negative scores indicate an event that shows displeasure by the users.

The system architecture is shown in Fig.1 and consists of five main blocks, Financial Community, the Tweets Filter, Sentiment Analyzer, Normalization and the Detect Events. Initially it is defined a financial community chosen by us from the universe of Twitter accounts which will be used to extract tweets for analysis. Therefore the collected tweets entering a filter with the aim of debugging only the tweets related to a specific stock or index like the Dow Jones Average. As the next step, with the tweets organized into files, our system will use the Sentiment Analyzer block for calculating the daily score of each tweet for every mood tool used. In the final stage, the normalization is performed to the files in order to normalize the volume of tweets. This normalization is required because the influx of tweets has been increasing over time. The last module of the system is event detection, that is, a module which is concerned with analyzing the scores calculated above to detect the occurrence of special events depending on the graph peaks. High peaks correspond to high scores that directly correspond to the occurrence of events that satisfy the users. In turn, low peaks correspond to negative scores that are equivalent to the occurrence of events that have sparked displeasure in the financial community.

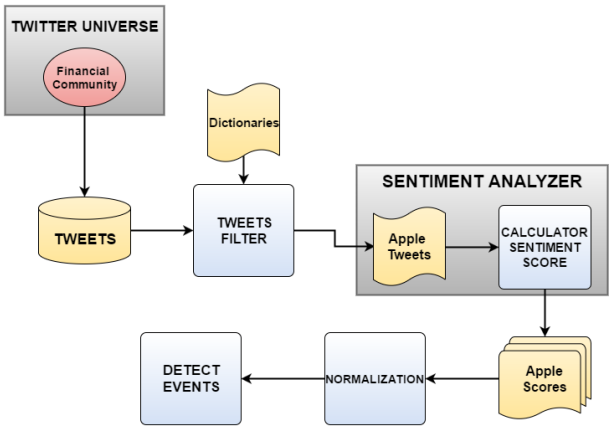

Fig. 1 - System architecture.

3.1 TWITTER

With over 600 million users, Twitter is a social network that allows its users to post quick text messages limited to 140 characters, known as tweets. Twitter is structured with followers and followed, where each user can choose who you want to follow and be followed. There is also the possibility to send messages privately to other profiles. It also allows, release videos, photos and directs the reader to other web pages through links. It is a dynamic social network that enables any user to have access to information that is constantly published. Contrary to Facebook where a user needs to ask to be friends and the friend needs to accept, here any user can follow another person without restrictions. According to the conceptions of (Recuero & Zago, 2009) Twitter allows users to create a public profile, interacting in this way with others through the published messages. In addition, you can show your network, offering ways to generate and maintain social values between these connections. It is also possible for the user to make some customizations, such as changing the background image, colors, and fill in personal data.

Twitter provides a (REST APIs, s.d.) for developers that allow them access to updates, status, data and users, among other features. REST (Representation State Transfer) is a style of network architecture hybrid derived from the style of various network-based architectures, which defines a connector interface that allows customers to "talk" to one-way servers. Twitter also provides developers access to a large volume of information in real time via an (Streaming APIs, s.d.). The tweets can be grouped by hashtag words preceded by the # character, used to mark keywords or topics in a tweet. In addition, users can make retweet in a tweet posted by other users. Twitter was used as data mining source for this work, considering that Twitter is a social network with a large number of active users. It is also a social network with global reach, which leads its users to share their ideas constantly. This sharing creates lot of information at every moment, and the APIs provide the access to this data tweets easily.

A) Definition of a Financial Community

We chose the social network Twitter as a source of data due to be one of the main online sources providing comments and discussions on the financial market. Using the Twitter API we implemented data collection software for the Tweets of a financial community. The financial community is composed of a subset of Twitter users with similar interests in the financial market. The objective of establishing this

community allows us to extract the most relevant Twitter users universe that somehow are related to the financial market.

The idea of the financial community is to choose users that are interested in the stock market and have more followers that people that they follow so they are able to change the market. We used 10 people very known as a seed to find these persons. In really it is not the 10 persons that are very important that we want to follow because what they say everyone knows because they appear on television, Bloomberg and Cnbc. It is the other power users not so know, but still with a big community of followers that we are interested in see what they are saying.

B) Construction of the Financial Community

The financial community proposed in this work begins with a collection of ten user accounts representing investment experts, financial news providers, managers and founders of companies. The selection criteria for these ten users is based in ten persons that are very known in the financial community, recommended to us by persons that work in trading stocks. But the final objective is to find the persons that follow them. The second stage of selection of user accounts was based on the followers of this community of ten users. There is a strong likelihood of these followers belong to the same financial investment community thus sharing similar interests. Yet there was a concern about the selected followers. Only the second stage users who have more followers than users that they follow were chosen. In Fig.2 is shown an illustration of this selection of the financial community. Basically the software developed accesses 10 influential users individually and their thousands of followers. For each follower the algorithm analyzes the number of followers and the number of people he follows. If the user has more followers than people he follows and the followers are more than 100 users it is concluded that this user enters in the financial community, because is an influential user of level two. For example in Fig.2, Ray Dalio user is one of 10 influential users in the financial market and has 6584 followers. Our software access all 6584 users and for each user the algorithm analyzes the followers. Users who have more people following than followers are discarded and are not included in the financial community.

After this filtering, we received the final financial community consisting of thousands of Ids. These Ids are the starting point for tweets that will be analyzed for use in the special events detection.

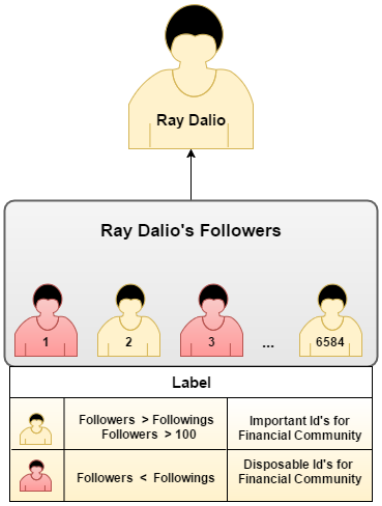

Fig. 2 - Example of a Financial Community Subset.

3.2 TWEETS FILTER

At this stage we have a tweets file for each user of our financial community.

The filtering is illustrated in Figure 3. With the database made, there was the need for a filtering of all content. The first filter makes use of a dictionary prepared by us containing 865 words related to the financial market and the stock exchange. The dictionary here proposed was based on a financial dictionary of (INVESTOPEDIA, 2016), where, in certain cases, we need to add some verb conjugations for relevant words. This dictionary contains the information that comes from the world of banking and investing, providing users with thorough and reliable meanings to all the most common, and even uncommon, financial terms. The objective of this initial filter is to delete all the tweets that have no content related to the financial market. Basically tweets passing to the next filtering stage must contain at least one of the words constituting the dictionary (Financial Dictionary. txt). For example, some words that the Financial Dictionary contains are: *asset, average, business, buy, capital, investment, gain, cash, cartel, market, ratio, crash, debt, dividend, economic, exchange, inflation, outsourcing, etc.* In the case of the word "buy" the dictionary has words with the conjugation of verbs "buy, bought, buying". For example, if a tweet that contains the word "apple" but that has nothing to do with the company, then as the word should not be linked to a tweet that contains financial content this tweet is discarded. That is, the tweet with the word "apple" that is related to the apple fruit does not pass the first filter and is eliminated. Basically it is assumed that a tweet that concerns the Apple Company contains financial content.

Then another filter has been applied in order to filter the tweets through a file with keywords related to the companies of the Dow Jones index. This filter is based on a file that contains a few keywords of each company. In Table 2 is shown an example of keywords for four of the thirty companies. Depending on the keywords that each tweet has the filter organizes tweets in thirty files where each file corresponds to one of thirty companies in the Dow Jones Industrial Average. For a tweet to match one of the companies it just takes one of the keywords listed in the file of keywords. The file consists of three types of keywords. The first is the company name, the second identifies the stock ticker which is characterized by the symbol '$' followed by its ticker symbol or stock symbol. Twitter created the “cashtag” as a way for people to create tweets about the stock market. The “cashtag” is the $ symbol plus the one to five letter

that are the stock symbol. Stock symbols are unique identifiers assigned to each security traded on a particular market. For example, $AAPL is for Apple Company. The last type of keyword is the hashtag that is composed of the '#' symbol followed by the company name.

Table 2 - Example of keywords used in the second stage of the filter for four companies of the DJIA.

COMPANY NAME	KEYWORDS		
Apple	APPLE	$AAPL	#APPLE
Boeing	BOEING	$BA	#BOEING
Cisco	CISCO	$CSCO	#CISCO
Nike	NIKE	$NIKE	#NIKE

For example, after the tweet "*Fri Sep 06 21:12:21 BST 2013 - Big news: $AAPL reportedly struck deal w/China Mobile to sell new cheaper iPhone to its 700 mil users*" passes the filters, once it contains the keyword *$AAPL*, it is placed in *Apple.txt* file.

At the end of these two steps of filtering we get thirty files, each file corresponding to a company, which contains all the tweets from the financial community about the company and relevant content on the financial market.

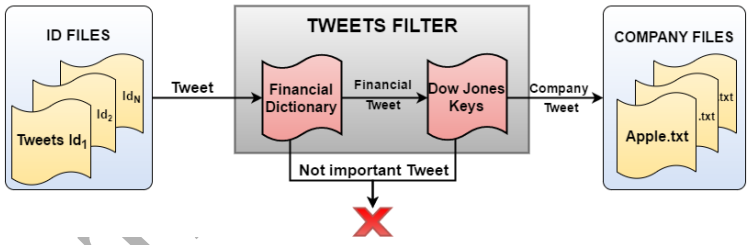

Fig. 3 - Filtering Tweets.

3.3 DEFINING SENTIMENT

It is often unclear whether a tweet contains some feeling. In this study, we used four text analysis tools to assess this feeling often difficult to find. One of the tools used was developed by us and the rest were developed by different authors. Next are described briefly and in Table 2 is an example of the evaluation of six tweets to four different tools.

- **MySentimentApi**

Our application to analyze the sentiment expressed in tweets was developed in Java and is also based on dictionary words that contain the polarity of words. Sentiment140 is a Web application that classifies tweets according to their polarity. The evaluation is performed using the distant supervision approach proposed by (Go, Bhayani, & Huang, 2009) that was previously discussed in the related work section. Sentiment140 Lexicon (version 0.1) it was created using a sample of 1.6 million tweets, and emoticons were used as positive and negative labels

The dictionaries used in our application were developed by the project sentiment140 and contain two lists, one consisting of unigrams and another for bigrams. The unigrams list comprises 62,468 terms (one term is one word) and the score respectively. The bigram list consists of sets of two words and the respective score in this set. The bigrams list comprises 677,698 terms (one term is the set of two words).

Our tool is developed based on the evaluation made in these dictionaries and basically our algorithm reads the tweet with the aim of adding scores taking into account words that they find. It should be mentioned that the algorithm implemented have the text segmentation that is the process that converts free text in single units with meaning. The text segmentation can be divided into: 1) Segmentation of words and 2) Segmentation of sentences.

As our data analyzed are tweets and usually have only one sentence (140 characters maximum) we only care about the part of 1) Segmentation of words. Segmentation of words is responsible for dividing the string finding the words division boundaries. The segmentation of words is often referred to as Tokenization, which means dividing a set of characters into tokens. In most European languages the delimiter tokens is space. However, many languages, such as Japanese and Chinese, do not use the space as a delimiter. Thus, tokenization is divided into two approaches: for languages in which the space is the delimiter, and for the token delimiter is not the space. In this algorithm the delimiter implemented is the space. Defined the delimiter token, our algorithm runs the tweet and whenever it encounters a word (a word is a string that are between spaces) analyzes the word taking into account the dictionaries. Initially, in our algorithm, the tweet is read and are added all the unigrams scores found in the tweet. In a second phase the tweet is re-read and are added all bigrams scores found in the tweet. The final result is the sum of the scores of unigrams and bigrams found in the tweet.

As previously mentioned, the sentiment analysis tool developed by us was based on two lexicons that are available on the Internet derived from other studies. The fact that we developed our tool based on these dictionaries with numerical polarities assigned to words presents a limitation in the event of detection. This limitation is visible on the accuracy of the tool compared the other as can be seen with the study cited in the state of the art, section 2.4, (Fernández-Gavilanes, Álvarez-López, Juncal-

Martínez, Costa-Montenegro, & González-Castaño, 2016) presents a new approach that applies a context-based algorithm to automatically create the dictionaries for each particular context.

- **TextBlob**

(Loria, s.d.) is a text analysis tool developed in Python can be used to perform various natural language processing tasks such as marking of part-of-speech, noun phrase extraction, sentiment analysis, text translation, and many more. TextBlob aims to provide access to word processing operations common through a familiar interface. The feeling property returns the sentiment in the form (polarity, subjectivity). The score polarity is a float in the range [-1.0, 1.0] and subjectivity varies within the range [0.0, 1.0], where 0.0 is very objective and 1.0 is very subjective.

- **Sentistrength**

This tool (SentiStrength, s.d.) makes use of a sentiment classification using unsupervised learning and is open source written in the Python language. This tool is a lexicon-based sentiment evaluator that is specially focused on short social web texts written in English. The classification will be in five different classes: positive, negative, neutral, extremely negative and extremely positive. As the tool is based on unsupervised learning, it makes use of a dictionary of positive and negative words. Different values are assigned to these positive and negative words, and the classification is based on how many positive and negative words appear in the sentence. For each passage to be evaluated, the method returns a positive score, from 1 (not positive) to 5 (extremely positive), a negative score from -1 (not negative) to -5 (extremely negative), and a neutral label taking the values: -1 (negative), 0 (neutral), and 1 (positive).

- **Affin**

(Nielsen, s.d.) it is an open source word processing library written in Python based on the Affective Norms for English Words lexicon (ANEW). Inspired in ANEW, the words were manually written by Finn Årup Nielsen in 2009-2011 based on an analysis of sentiment in short texts found in social life and in the media. Positive words are scored from 1 to 5 and negative words from -1 to -5, reason why this lexicon is useful for strength estimation. The lexicon comprises 2477 words in English (including some sentences) with the respective sentiment evaluation.

In Table 3 an example of six tweets is presented with the appropriate calculation of scores for the four sentiment applications mentioned above. The table below shows a normalization of values in a range of -5 to 5 to be easier to understand and compare the four tools of sentiment analysis.

As can be seen, the content of the first two tweets theoretically are positive. And in practice, the four tools calculate a positive score. The same applies to the two negative tweets presented in the following lines, and it is easy to see that the tweets are negative and the four tools also classify them as negative. The difficulty found in the tools is to differentiate positive tweets neutral tweets. As can be observed, the last two tweets are composed of neutral content. Both TextBlob and Affin are able to calculate a neutral score. Instead the application developed by us thinks it is positive score and SentiStrength tool too. Our

application even has a small margin of error compared to positive tweets displayed above. But on the contrary Sentistrength presents very similar ranges. That is, this Sentistrength tool presents a greater error in the detection of positive and neutral tweets.

Table 3 - Example sentiment evaluation tweets.

THEORETICAL ANALYSIS	TWEETS	MYSENTIMENT API	TEXT BLOB	SENTI STRENGTH	AFFIN
Positive	"This bigger pixels thing by Apple makes HTC look good. Very good."	**3,5**	**3,3**	**2,8**	**4,1**
Positive	"We live in an exciting period of transformation in #Technology - #MicrosoftHoloLens"	**4,3**	**3,5**	**2,7**	**3,6**
Negative	"#Apple stock is falling since today's announcement. Not surprised."	**-0,5**	**-1**	**-1**	**-1**
Negative	"Cisco is down 12% after a disappointing earnings release. @stocksnboxing tells CNBC why he thinks it's still a buy: http://t.co/7h3EZVs7Ok"	**-4,7**	**-1,5**	**-1,8**	**-2,3**
Neutral	"Walmart informational meeting in GB over plans for downtown @CityofGreenBay @WFRVNews #walmart http://t.co/5ywrgJ2R72"	**2,4**	**0**	**2,1**	**0**
Neutral	"Catch me on CNN tomorrow at 6:40AM talking Apple's announcements. Tune In!"	**2,2**	**0**	**2,1**	**0**

3.4 NORMALIZATION

Normalization of data is an important step in the system architecture. Since the use of the Twitter social network has been increasing over time it is important to normalize all the data in relation to the time when the tweet was created so that there is consistency in relation to the volume of tweets over the period of the simulation. As can be seen in Fig. 4 it is clear that the number of tweets increases exponentially between 2013 and 2015. Through this graph is easy to see that it is necessary to make a normalization volume of tweets, so as to achieve conformity in very positive and negative peaks. Having said this was implemented a daily sliding window with the average volume of tweets in the last three months. The following formula (1) was applied to all values of the files for each of the four sentiment analysis tools.

$$Score_{Normalized} = \frac{Score_{daily}}{\frac{Total\,Tweets_{last\ 3\ months}}{90}} \quad (1)$$

Normalization for each day is to divide the daily score by the average volume of tweets in the last three months. This normalization makes everyday theoretically have the same volume tweets. Fig. 5 shows two graphs with an example for Apple Company to better understand the effect of normalization in the data. As shown in Fig. 5, the normalization was successful since the special Apple event which marked the launch of the iPhone 6 was more important than the launch of the iPhone 6S. That is, before normalization, the day of the launch of the iPhone 6s had a higher volume of tweets compared to the launch day of the iPhone 6, which was reflected in the final score. After normalizing by the volume, ie after getting a constant volume of tweets over time it is concluded that the score is more positive on the day of launch of the iPhone 6 than on launch day iPhone 6s thus proving the theory.

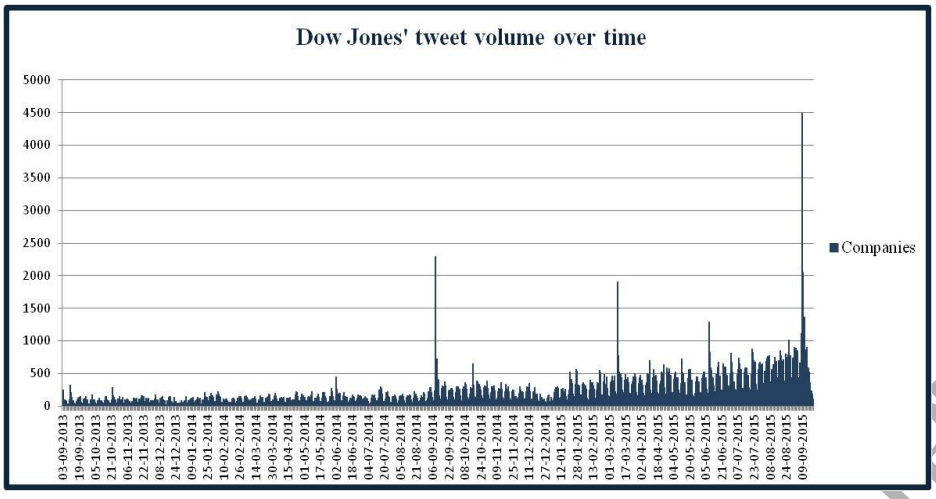

Fig. 4 -Total Tweets over time for Dow Jones stocks.

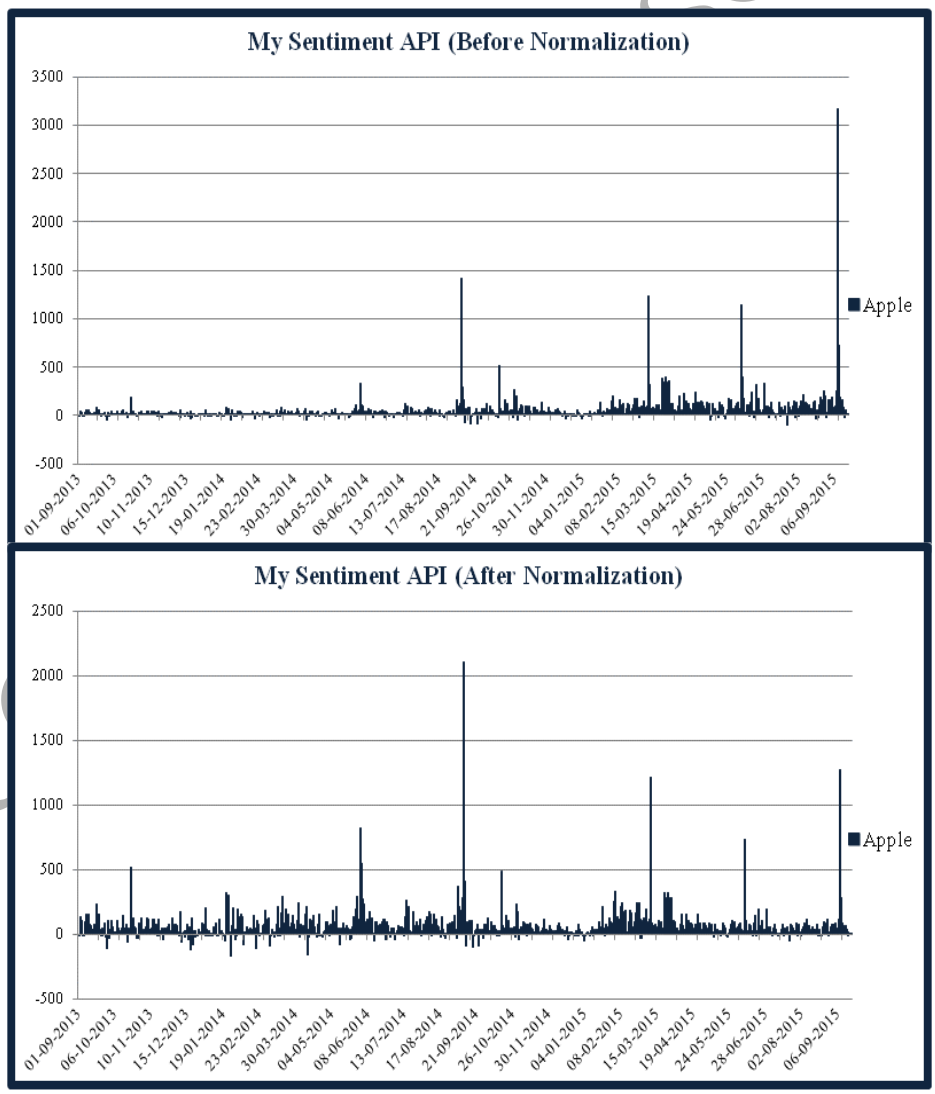

Fig. 5 - Example normalization Apple's Company.

3.5 Detection of Events

This module is responsible for detecting events during the company's test period, where events are here identified as very negative or very positive peaks in the time series. These peaks result from the analysis of the daily sentiment time series extracted from tweets published concerning the company. The analysis of sentiment as stated earlier, is carried out with four different tools, hence resulting in the final four graphs for each company with the respective developments. This module is responsible for analyzing the four graphs in parallel and see if there is a consistency in the peaks with the respective events.

The algorithm to detect the events is quite simple, that is the reason why it was not expressed in a formula, it results from a "score" value, which only needs to be larger than a certain score threshold. This event detection module has the objective of being an indicator to be used as a feed to a genetic algorithm software. In the case of implementing an artificial intelligence algorithms (Genetic Algorithms) using the event detection module as an indicator, is the algorithm that is responsible for deciding the value assigned to the score level at which it detects an event. But it will be easy to view that per year we only have 5 to 10 important events. The others that are too small are not important since they do not change the stock market, so we are not interested in them.

4. Results

For the next experiments we use the tweets from two years to detect special events for each of the companies studied. All tweets are from more than 9000 users who follow the 10 user accounts set initially. The extracted data is from September 2013 to September 2015.

4.1 Pre Processing data collected

In total, we extract more than 12 million tweets of 9011 user accounts. These tweets have produced a first filtering for only the tweets that contain content relevant to the financial market. This filtering was performed using a dictionary prepared by us containing hundreds of words on finance and market. In a next step the filtration was also performed based on a dictionary with the goal of creating thirty files, each representing a company that belongs to the Dow Jones Average. This dictionary also developed by us has some keywords for each company, such as the name of the company, the company symbol, the hashtag, and others. At the end of these filters we have obtained a total of 192,935 tweets for analysis. Table 3 shows some relevant information in the data collection stage to the completion of the project.

Table 3 - Information on the stage of collecting tweets.

Description	Quantity
Number of days testing	710
Influential users	11
Financial community users	9011
Tweets collected	12.328.766
Tweets filtered	192.935
Total size on disk tweets collected	1,61 GB
Total size on disk tweets filtered	27,1 MB

4.2 DEVELOPMENT AND VALIDATION OF CLASSIFICATION MODEL

After the processing of the collected tweets the next step is the classification of the implicit sentiment in the tweets. Each tweet has been classified by the four text tools presented before.

A) APPLE – CASE STUDY

The first case study presented refers to the Apple Company. Apple is an American multinational corporation that aims to design and market consumer electronics, computer software, and personal computers. The company is known for its special events, which serve to announce new products, new product designs and improvements through press conferences that bring together a significant number of followers. Many times the purpose of the event is kept secret to trigger curiosity and noise from the users and the purpose of the event is only revealed during the event. In Fig. 6 are shown four graphs for each Humor tool used to analyze the sentiment expressed in tweets related to the company in question, as it was explained in the previous section. Each graph shows the evolution of the company over the tweets collection period. In each graph the special events were detected and numbered that are relevant in the company's life during the period.

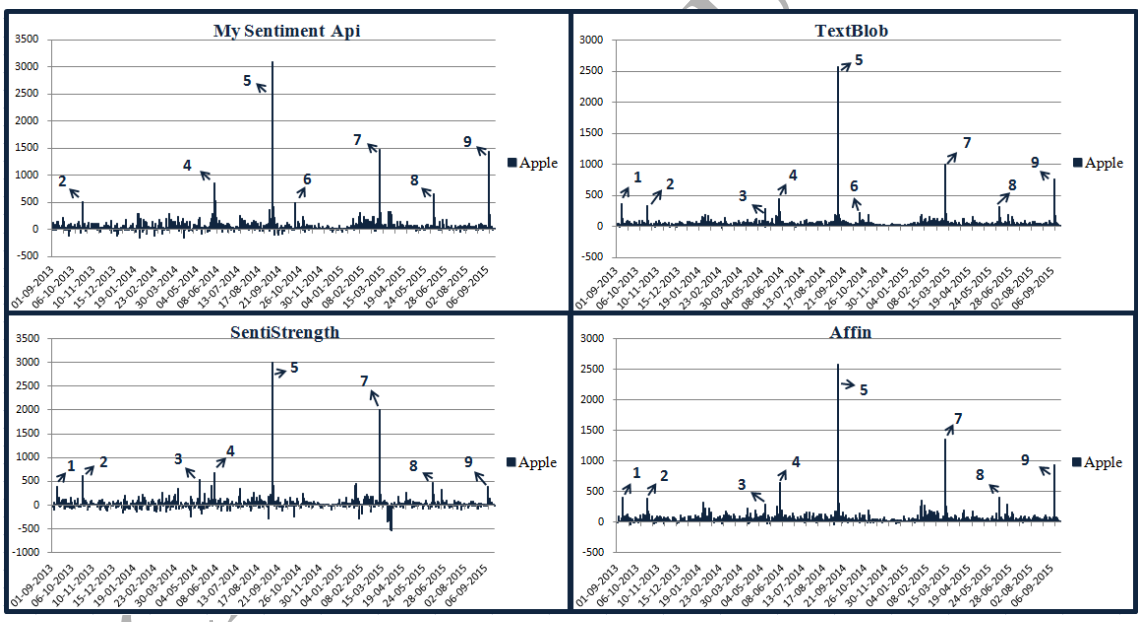

Fig.6 - Analysis sentiment for the four tools of humor over time – Apple's Company.

The special event for each number in this graph is indicated in Table 4. As can be seen, the peak number **5** (09-09-2014) is the one with greater consistency in the four graphs. It was a special event much talked by the financial community as it referred to the long-awaited release of iPhone6. Another special event well marked in the graphs is the peak number **7** (09-03-2015) which is characterized by the launch of Apple's Smart Watch. This event is already waiting for several years by the brand's followers because it has a small revolution that shows how the idea of Apple for Smart Watch, adjusted to the user and with all the necessary features. The special event number **9** (09-09-2015) also features quite excitement by the financial community although not as imposing as the event number 5 because the number 9 is just an upgrade of iPhone6, ie is an event characterized by the launch of the iPhone 6S. Below in Table 4 are

examples of tweets analyzed on days of events in order to confirm that users published information about the special events.

Table 4 – Examples of tweets for special Apple Event.

DATE (NUMBER)	APPLE SPECIAL EVENT	TWEETS
10-09-2013 (1)	**Apple announced the iPhone 5C and iPhone 5S**	"#Apple unveils more powerful fingerprint-scanning #iPhone 5S, multi-colored iPhone 5C http://t.co/KtEIZEZcOB via @TimesTech"
22-10-2013 (2)	**Apple announced iPad Air and iPad Mini**	"RT @Forbes: The new iPad Air and iPad Mini will maintain Apple's premium positioning in the tablet market http://t.co/4uWsNW4I9J"
09-05-2014 (3)	**News: Apple reveals iPhone 6 in August, the newspaper says**	"RT @FirstReporter24: Apple iPhone 6 got a launch date finally. Launching in August, 2014! #technews #tech #technology #Apple"
02-06-2014 (4)	**Apple presented the new version of OS X and iOS.**	"Apple announces iOS 8. Tim Cook calls it giant release."
09-09-2014 (5)	**iPhone6 and iPhone6 Plus**	"#Apple live blogging and live streaming #iphone6 event in 1 hours http://t.co/92hYo1Ln6F"
		"Apple's new iPhones are to be called the #iPhone6 and the iPhone 6 Plus. #AppleEvent"
		"#Apple #iPhone6 has 1 million pixels and #iPhone6plus 2 million pixels... Both thinner than any iPhones ever made. @NBCLA"
		"Sweet! The iPhone6 and iPhone6 Plus looks great! #AppleLive"
		"RT @mashabletech: #AppleLive: #iPhone6 has a rounded, seamless surface that Apple says is created using "precision polishing process." "
09-03-2015 (7)	**Smart Watch Apple**	"$AAPL up 0.8% pre-market ahead of Watch event. ETA 5 hours"
		"Lots of interesting Apple Watch speculation from @daringfireball - fun prep for the big event! http://t.co/iu2EqJoNHf http://t.co/SoQoeu1Ju4"
		"Cook: The Apple Watch is the most advanced timepiece ever created."
		"Arghhhhh! Get on to the Apple watch already!!"
		"I don't want the Apple Watch."
08-06-2015 (8)	Apple launches the big WWDC	"Happening now. Apple's #WWDC2015 Christmas in June. Boom.http://t.co/TI8p98gv5b http://t.co/R5hntDmtQb"
09-09-2015 (9)	**iPhone 6s, 6s Plus, Apple TV e iPad Pro**	"New #iPhone6s and #iPhone6sPlus are "the most advanced smartphones in the world." #AppleEvent"
		"Sign me up! NEW iPad Pro that is 1.8x faster than iPad Air 2, new keyboard & Apple pen #sweet. #AppleEvent http://t.co/j2sp2myQel"
		"Nice camera on the new #iPhone6s. #AppleEvent"
		"All the deets #iPhone6s #iPhone6sPlus Specs, pricing, availability & Apple's new iPhone Upgrade Program #AppleEvent http://t.co/XGhkb1afOM"
		"RT @AboveAverage: BUT WILL THE NEW APPLE TV HAVE A STOCKS APP???????#wewantstocks #AppleEvent #stocks"
		"@tim_cook: #iPhone6s #iPhone6 "the most loved phones in the world." #AppleEvent"

From the previous tweets it is possible to observe in Table 4 that in the days of the special events the tweets have mainly positive or very positive sentiment. For the remaining peaks of the graph (1, 2, 3, 4, 6 and 8) that are numbered, they have less impact than those mentioned above. But these peaks also have high interest from the users as they have more positive tweets compared to the remaining days. As described in Table 4, we have the example of the event where Apple launches iPhone5C, events in which Apple has updates on models. There is also event 3 which is the day when it was announced the supposed launch of the iPhone6. This news created agitation on Twitter as the event in which Apple released iPhone6 was the most significant event during this review period.

B) MICROSOFT – CASE STUDY

The second case study concerns the Microsoft Company. The Microsoft is an American multinational corporation that develops, manufactures, licenses, sells and supports computer software, electronic products, computers and personal services. Among its best known software products are the lines of Windows operating systems, applications line for Office and Internet Explorer browser. In the following Fig. 7, shows the graphs for each sentiment tool used.

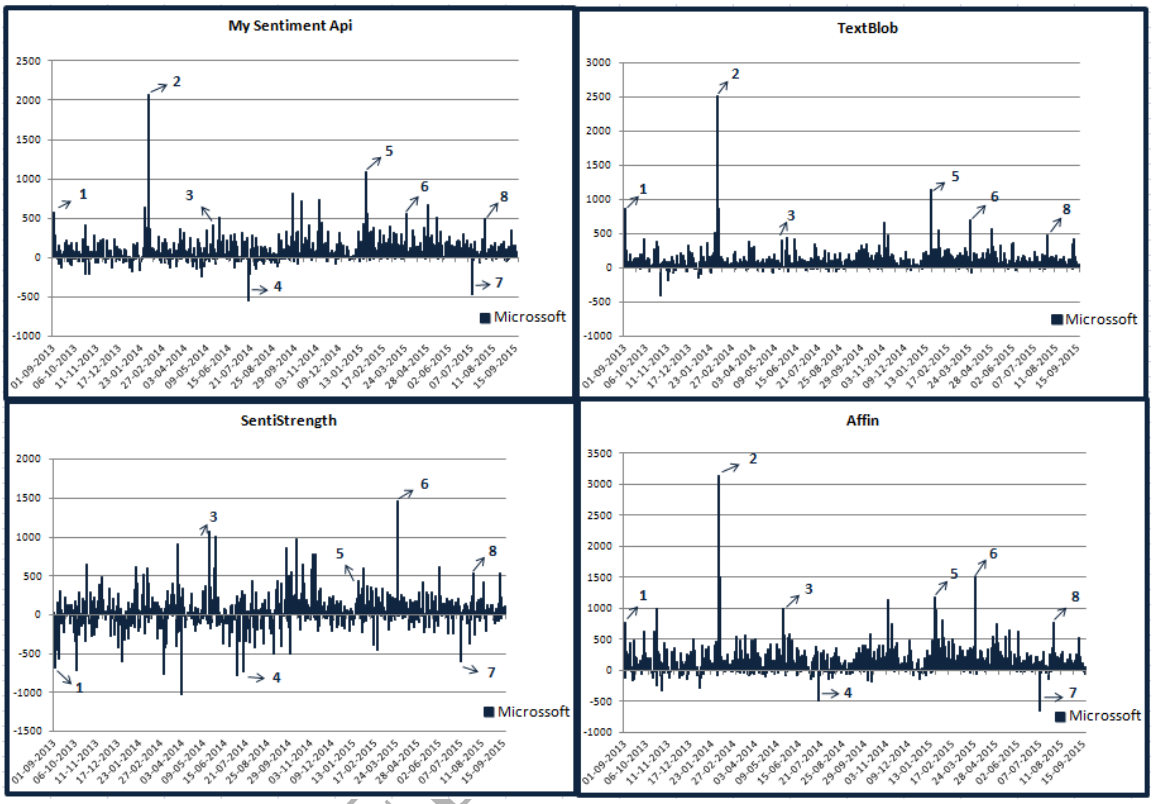

Fig.7 - Analysis sentiment for the four tools of humor over time – Microsoft's Company.

The first event highlighted in the Fig. 7, it is the number 1 event and concerns the announcement of Nokia's purchase by the Microsoft Company. On 03-09-2013 Microsoft announced the purchase of Nokia, in particular, the segment of mobile services and devices. With this acquisition, Microsoft aims to make even stronger Windows Phone and face competitors Apple and Google. The number 1 event showed satisfaction by the users when the evaluation is performed by the tools MySentiment Api, TextBlob and Affin. Already SentiStrength tool application displays a negative score respectively to this day. We can see through the Table 5 examples of tweets published in the days of the events highlighted in the graphs.

The most remarkable event in this case study is the number 2 event. Is an event that attracted much agitation by users because it is related to the presentation of the new Microsoft CEO, Satya Nadella. It can be confirmed in Table 5 examples of tweets that show the publication of content about this presentation of the Microsoft CEO.

Another day that was not indifferent to the users for the worst reasons was the 17-07-2014. It was a day marked by the discontent of users who follow the brand. This event, number 4, is characterized by the announcement that the company made regarding the largest wave of dismissals ever. The company made a cut in more than 18,000 employees. The event number 7 also demonstrates the same displeasure that the event 4. As well as the event 4 in the event 7 were also cut over 7,800 jobs. As we can confirm with the evaluation scoring in the graphs this events generated many negative comments from users.

As mentioned, one of the most known products of the company is the lines of Windows operating systems. An event that attracted interest was the number 5 event which served to promote the new version of the Windows operating system (Windows 10), which was released later day 07-29-2015, event number 8. In addition to software products the company is also equipped with hardware products. Between Xbox video game consoles, the series of Surface tablets (Event 3) and Smartphones Microsoft Lumia, Nokia old (03-09-2013, first event).

Table 5 – Description of Microsoft's Special Events with examples of tweets

DATE (NUMBER)	MICROSOFT SPECIAL EVENT	TWEETS
03-09-2013 (1)	**Microsoft buys Nokia**	"RT @Microsoft: Microsoft to acquire Nokia Devices & Services: http://t.co/wZHhgLvvyJ"
		"After 2 yrs of cutting through all their cash reserves, Nokia is finally sold to Microsoft. I wonder what's next. Blackberry looks doomed!"
		"RT @yazanalsaeed: Microsoft to acquire Nokia's handset business for $7.2 billion http://t.co/HALNMOmAPz #Microsoft #Nokia #digital"
04-02-2014 (2)	**Microsoft announces new CEO, Satya Nadella**	"RT @Microsoft: Introducing our new CEO, Satya Nadella: http://t.co/u5IGl1N78G"
		"Microsoft's new CEO brings 22 years of experience - of Microsoft"
		"Bill Gates quits as Microsoft chairman as Satya Nadella is named chief executive | via @Telegraph http://t.co/LnQtO5rECf."
20-05-2014 (3)	**Microsoft Surface Event**	"Waiting on @Microsoft Surface event to start. http://t.co/c0eRFLF2SI"
17-07-2014 (4)	**Microsoft cuts 18,000 jobs**	"'Large Round of Layoffs Expected at Microsoft" by NICK WINGFIELD via NYT"
		"Microsoft eliminating up to 18,000 jobs. http://t.co/atcuqCYjDc"
		"18,000 job cuts at $MSFT represent 14% of the workforce and the pretax charge of ~$1.5B equates to 6.5% of estimated 2014 profits. Painful."
21-01-2015 (5)	**Event to promote Windows 10**	"Join CNET for live coverage of Microsoft's Windows 10 event! http://t.co/c5osAipdLJ via @CNET"
26-03-2015 (6)	**Microsoft launches Surface Pro 3**	"Microsoft makes cheaper version of Surface Pro 3, with smaller screen, less-flexible kickstand: Microsoft is m... http://t.co/Mm6a98vZzd"
08-07-2015 (7)	**Microsoft cuts 7,800 jobs**	"RT @NEWS1130: Microsoft to cut up to 7,800 jobs, mostly in phone hardware. It expects an impairment charge of about US$7.6 bln"
29-07-2015 (8)	**Microsoft launches Windows 10**	"Windows 10 @microsoftgulf launches today. How is your desktop looking ? http://t.co/lhehtYf4lB http://t.co/czsKrBpHFn"

To conclude this case study is important to note that sometimes there is a lack of consistency in the four graphs although able to detect very well certain events in the company. The events highlighted correspond to the great events that have occurred in the company during the period studied. As can be seen in the table above, Table 5, the existing tweets in our data collection correspond to the detected content in the events.

C) WALMART – CASE STUDY

Finally, the third case study analyzed relates to the Walmart Company. Walmart is an American multinational retail corporation that operates a chain of hypermarkets, discount department stores and grocery stores. Is the world's largest company by revenue, according to the Fortune Global 500 list in 2014, as well as the biggest private employer in the world with 2.2 million employees. Fig. 8 shows the graphs corresponding to the sentiment evaluation of the company for each of the four tools used in the test period.

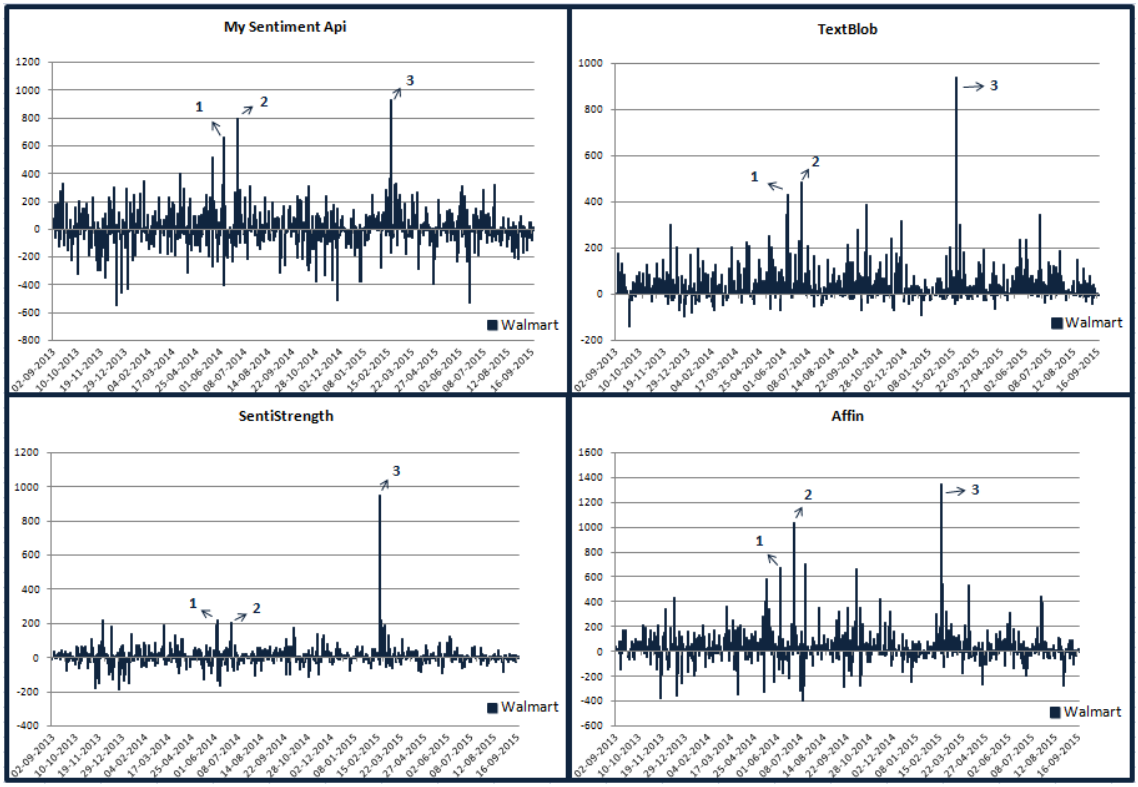

Fig.8 - Analysis sentiment for the four tools of humor over time – Walmart's Company.

In this case study there are only three main events. The most notorious event is registered with the number 3. This event detected on February 19, 2015 was marked by the announcement made by the Walmart on the increase of wages of its employees. In a statement Walmart pledged to increase the salary of 500,000 employees. All employees will receive a pay rise to at least US$ 9 per hour. In February next year, the value per hour increases to US$ 10. The announcement was made with the release of results for the fourth quarter of 2014. This announcement was received with great satisfaction by the users being reflected in the tweets published that day as shown in Table 6.

Another event that has not been indifferent to the financial community proposed here was the first event detected in Fig. 8. The Annual Shareholders' Meeting is an annual event held on the day 6 June 2014. In Table 6 it can be seen two examples of published tweets demonstrating interest in the event that occurred on that day.

As discussed in the first case study presented in this section the launch of Apple's devices are very important events in the life of the financial community. As we can see on the day of event September 10,

2013 the company Apple launched the iPhone 5C and iPhone 5S. In an announcement on Thursday (June 26, 2014, second peak), big-box retailer Walmart said it will cut the price on the iPhone 5C and iPhone 5S, starting at 09:00 on Friday. The fact that the company reduced prices, generated a wave of satisfaction of users who saw a great opportunity to get the Apple devices at a better price. This day it was also said that the move to cut the prices of iPhones may fuel speculation that Apple (AAPL, Tech30) is planning a new version of the iPhone. Retailers often cut prices before new releases to clear inventory. As was seen in the first case study for the company Apple, around two months after the Walmart event number 2 the speculation was confirmed when Apple launched the iPhone 6 and iPhone 6 Plus (September 9, 2014). Table 6 presents two examples of tweets that reflect the previously explained. The first refer the low prices of the iPhones on Friday and the second tweet also shows a bit of speculation around the launch of the iPhone 6.

Table 6 – Description of Walmart's Special Events with examples of tweets.

DATE (NUMBER)	WALMART SPECIAL EVENT	TWEETS
06-06-2014 (1)	**Announces 2014 Annual Shareholders' Meeting Voting Results**	"Walmart's shareholder meeting is today. Some things they should be talking about: http://t.co/Hj0tlLTBl4 #WalmartEconomy"
		"RT @JillianBerman: Pharrell on stage at Walmart shareholders meeting "make some noise for Walmart" http://t.co/Yvw0lTW66y"
26-06-2014 (2)	**Drop the price of iPhones**	"Walmart to cut iPhone 5C and 5S prices on Friday http://t.co/w4lfzjGkdm"
		"Walmart is dropping the price of its iPhone 5s to $99. This is a good sign the iPhone 6 is around the corner. http://t.co/l0xkDitUJN"
19-02-2015 (3)	**Increase the wages of employees**	"Another beneficiary of @Walmart boosting hourly wages is Walmart. Its hourly workers are some of its best customers. $WMT"
		"HIGHER pay = better service = happier customers = better results = higher stock. @Walmart CEO thinking on boosting pay of 500k associates."
		"Good move! "@FinancialTimes: Walmart to raise pay of 500,000 employees http://t.co/XpO3KjE4CN "

This last case study was a case study with fewer detected events Compared to the last two. Although one of detected peaks allowed the relationship with the first case study, managing to prove satisfactory event detection.

5. CONCLUSIONS

This paper focuses on event popularity through sentiment analysis of the tweets posted by a defined financial community. This paper proposes four different text analysis tools as a basis for evaluating the content of the extracted tweets from September 2013 to September 2015. The presented work provides a basis for event popularity that has to deal with massive amounts of data and very high levels of noise typical of social networks. The model handles a large amount of data and detects events efficiently. It has achieved consistency between the four sentiment tools getting most of the time the same events. The results show the good performance of our model demonstrating that defined financial community is influential with regard to the publication of tweets about companies, and able to detect financial events.

There are several ways in which this work can be extended:

Improve the creation of the financial community defined by us. It is important to establish a strong financial community. In this work we realized that the implemented financial community was very useful

for completing the work as these tweets is the base for the rest of the work. The truth is that today there are various applications and techniques that allows a rogue user to increase the number of followers on social networks. That can decrease the quality of the users chosen by the algorithm.

Explore the importance of retweets in the event detection process. A retweet replicates something that was written by another user. This happens when a user writes a sentence about something that is interesting to others, or when it is a matter of public interest, which must be passed forward.

Dealing with noise could be addressed from the perspective of supervised learning, where the main challenges are in choosing an efficient and incremental learning algorithm using a minimal amount of training data and addressing the potential need for retraining.

Test different Financial Communities by starting with different user accounts. See if the results are similar or different based on different seeds of user accounts.

Finally we could combine sentiment analysis tools to the public tool Google Trends. This tool shows how many times a particular search or term is entered relative to the total search-volume in several regions of the world and in various languages. This would be useful tool to introduce to this type of work to be able to understand the trend of a particular brand, product or subject over time.

Acknowledgment

This work was supported in part by Fundação para a Ciência e a Tecnologia (Project UID/EEA/50008/2013).

6. REFERENCES

Atefeh, F., & Khreich, W. (2015). A survey of techniques for event detection in twitter. *Computational Intelligence, 31 (1)*, 132-164.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). *SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining* (Vol. 10).

Balazs, J. A., & Velásquez, J. D. (2016). Opinion Mining and Information Fusion: A survey. *Information Fusion, 27*, 95-110.

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

Bradley, M. M., & Lang, P. J. (1999). Affective Norms for English Words (ANEW) Instruction Manual and Affective Ratings. *Technical Report C-1, The Center for Research in Psychophysiology.* University of Florida.

Brown, Stephen, J., Goetzmann, W., & Kumar, A. (1998). The Dow theory: William Peter Hamilton's track record reconsidered. *The Journal of finance, 53(4)*, 1311-1333. Retrieved from http://www.investopedia.com/university/dowtheory/

Chan, W. S. (2003). Stock price reaction to news and no-news: Drift and reversal after headlines. *Journal of Financial Economics, 70(2)*, 223-260.

Chen, H., De, P., Hu, Y., Jeffrey, & Hwang, B.-H. (2013). Customers as advisors: The role of social media in financial markets. *3rd Annual Behavioural Finance Conference.* Queen's University, Kingston, Canada.

Cunningham, L. A. (1997). The Essays of Warren Buffett: Lessons for Corporate America. *Cardozo Law Review, 19*, 1-220.

Deschatre, G. A. (2009). *Investimento em ações: Para os momentos de crise e de crescimento.* Rio de Janeiro: Thomas Nelson.

Esuli, A., & Sebastiani, F. (2006). *Proceedings of LREC. Sentiwordnet: A publicly available lexical resource for opinion mining* (Vol. 6). Citeseer.

Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance 25*, 383-470.

Fernández-Gavilanes, M., Álvarez-López, T., Juncal-Martínez, J., Costa-Montenegro, E., & González-Castaño, F. (2016). Unsupervised method for sentiment analysis in online texts. *Expert Systems with Applications, 58*, 57-75.

Geva, T., & Zahavi, J. (2014). Empirical evaluation of an automated intraday stock recommendation. *Decision Support Systems, 57*, 212-223.

Go, A., Bhayani, R., & Huang, L. (2009). Twitter Sentiment Classification using Distant Supervision. *CS224N Project Report, 1*, 12.

Gomide, C. S., Lima, A. A., Gomide, J. S., Roque, M. D., & Silva, S. T. (2014). *O Twitter como Instrumento de Detecção de Epidemias de Dengue e Desenvolvimento de.* Rio de Janeiro.

Gomide, C. S., Lima, A. A., Gomide, J. S., Roque, M. D., & Silva, S. T. (2014). O Twitter como Instrumento de Detecção de Epidemias de Dengue e Desenvolvimento de Políticas Públicas. *In: XXXVIII Encontro da ANPAD - EnANPAD.* Rio de Janeiro.

Gorgulho, A., Neves, R., & Horta, N. (2011). Applying a GA kernel on optimizing technical analysis rules for stock picking and portfolio composition. *Expert systems with Applications*, *38(11)*, pp. 14072-14085.

Haugen, R. A. (2001). *Modern investment theory.* New Jersey: Prentice Hall.

Hirshleifer, D. (2001). Investor Psychology and Asset Pricing. *The Journal of Finance*, 1533-1597.

Indurkhya, N., & Damerau, F. J. (2010). *Handbook of Natural Language Processing* (Vol. 2). CRC Press.

INVESTOPEDIA. (2016). Retrieved from http://www.investopedia.com/dictionary/

Jain, T. I., & Nemade, D. (2010). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. *International Journal of Computer Applications, 7(5)*, 12-21.

Lampos, V., & Cristianini, N. (2012). Nowcasting events from the social web with statistical learning. *ACM Transactions on Intelligent Systems and Technology (TIST), 3(4)*, 72.

Lampos, V., & Cristianini, N. (2012). Nowcasting events from the social web with statistical learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 72.

Loria, S. (n.d.). *TextBlob: Simplified Text Processing*. Retrieved January 2016, from https://textblob.readthedocs.org/en/dev/

Mendoza, M., Poblete, B., & Castillo, C. (2010). *Proceedings of the first workshop on social media analytics. Twitter Under Crisis: Can we trust what we RT?*

Mendoza, M., Poblete, B., & Castillo, C. (2010). *Proceedings of the first workshop on social media analytics: Twitter Under Crisis: Can we trust what we RT?*

Milagros, F. G., Tamara, Á. L., Jonathan, J. M., Enrique, C. M., & Francisco, J. G. (2016). Unsupervised method for sentiment analysis in online texts. *Expert Systems with Applications, 58*, 57-75.

Miller, G. A., Beckwith, R., Fellbaum, C., & Gross. (1990). Introduction to wordnet: An on-line lexical database. *International journal of lexicography, 3*, 235-244.

Nielsen, F. A. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*.

Nielsen, F. A. (n.d.). *AFINN: A new word list for sentiment analysis on Twitter*. Retrieved 2016, from https://finnaarupnielsen.wordpress.com/2011/03/16/afinn-a-new-word-list-for-sentiment-analysis/

Petrović, S., Osborne, M., & Lavrenko, V. (2010). Streaming first story detection with application to twitter. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, (pp. 181-189).

Recuero, R., & Zago, G. (2009). Em busca das" Redes que importam": Redes Sociais e Capital Social no Twitter. *XVIII Congresso da Compós, PUC/MG.* Belo Horizonte.

REST APIs. (n.d.). Retrieved 2015, from https://dev.twitter.com/rest/public

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). *Earthquake shakes Twitter users: real-time event detection by social sensors.* ACM.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). *Earthquake shakes Twitter users: real-time event detection by social sensors.*

Sankaranarayanan, Samet, H., Teitler, B., Leiberman, M., & Sperling, J. (2009). *Twitterstand: news in tweets.*

Sayyadi, H., Hurst, M., & Maykov, A. (2009). *ICWSM: Event Detection and Tracking in Social Streams.*

Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The AZFin text system. *ACM Transactions on Information Systems, 27(2)*, 12.

SentiStrength. (n.d.). Retrieved 2016, from http://sentistrength.wlv.ac.uk/

Silva, A., Neves, R., & Horta, N. (2015). A hybrid approach to portfolio composition based on fundamental and technical indicators. *Expert Systems with Applications, 42(4)*, 2036-2048.

Singal, V. (2006). *Beyond the Random Walk: A Guide to Stock Market Anomalies and Low-Risk Investing.* Oxford University Press on Demand.

Starbird, K., & Palen, L. (2010). Pass it on?: Retweeting in mass emergencies. *International Community on Information Systems for Crisis Response and Management.* Seattle, WA, USA.

Starbird, K., & Palen, L. (2010). Pass it on?: Retweeting in mass emergencies. *Information Systems for Crisis Response and Management Conference.* Seattle, WA, USA.

Streaming APIs. (n.d.). Retrieved 2015, from https://dev.twitter.com/streaming/overview

Taboada, M. (2016). Sentiment Analysis: An Overview from Linguistics. *Annual Review of Linguistics, 2*, 325-347.

Tetlock, P. C. (2007). Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *The Journal of Finance, 62(3)*, 1139-1168.

Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology, 63(1)*, 163-173.

Vega, C. (2006). Stock Price Reaction to Public and Private Information. *Journal of Financial Economics, 82(1)*, 103-133.

Wang, H., Can, D., Kazemzadeh, A., Bar, F., & Narayanan, S. (2012). *Proceedings of the ACL 2012 System Demonstrations - A system for real-time twitter sentiment analysis of 2012 us presidential election cycle.* Association for Computational Linguistics.

Winerman, L. (2009, January 21). Social networking: Crisis communication. *Nature, 457(7228)*, 376-378.

Winerman, L. (2009, January 21). Social networking: Crisis communication. *NATURE, 457*, 376-378.

Xiaodong, L., Haoran, X., Chen, L., Jianping, W., & Xiaotie, D. (2014). News impact on stock price return via sentiment analysis. *Knowledge-Based Systems, 69*, 14-23.

Yang, Steve Y.; Mo, Sheung Yin Kevin; Zhu, Xiaodi. (2014). An Empirical Study of the Financial Community Network on Twitter. *Computational Intelligence for Financial Engineering \& Economics (CIFEr), 2104 IEEE Conference on*, (pp. 55-62).