TÉCNICO LISBOA

Forecasting Stock Returns Based on Event Detection from Twitter

Mariana Gonçalves Alves Daniel

Thesis to obtain the Master of Science Degree in

Electronics Engineering

Supervisor: Prof. Rui Fuentecilla Maia Ferreira Neves

Co-Supervisor: Prof. Nuno Cavaco Gomes Horta

Examination Committee

Chairperson: Prof. Pedro Miguel Pinto Ramos

Supervisor: Prof. Rui Fuentecilla Maia Ferreira Neves

Members of the Committee: Prof. Pável Pereira Calado

May 2016

Acknowledgment

It is not possible to present this dissertation without first thanking all the people who in some way contributed to its development and execution.

First I would like to thank Instituto Superior Técnico for the opportunity to produce this thesis in order to expand my scientific background. In particular, I would like to thank the teachers Rui Fuentecilla Neves and Nuno Cavaco Horta, whose patience, help, advice and supervision were invaluable.

To my friends, work colleagues and employees of this institution, I thank you for all these years, for walking beside me and making this journey less difficult. I am grateful for the love and affection given throughout my academic career.

Last but not least, I would like to praise the people who are most important in my life. Without them, surely, I would not have finished my Master Degree with the motivation that I always had.

Thanks to my father, António Daniel, for the example of determination and spirit of sacrifice. No doubt you are my hero and thanks to your help I'm ready to conquer the world.

To my mother, Dina Daniel, by the inexhaustible patience and the ability to always have a kind word at the right time. You have always succeeded in attenuating the physical distance with simple phone calls at the most opportune moments, no doubt my emotional engine along this walk.

Thanks to my sister, Luisa Daniel, who never let me lose focus and always supported me in difficult moments remembering me always of the value I have. Thanks also to my brother in law, Alexandre Contente, and to my loving god daughter, Camila, who I have missed so much. Thank you for your love.

Because they make the world I live in a more special place, I dedicate this thesis to my family.

Resumo

O rápido crescimento de utilizadores no Twitter faz com que esta rede social seja uma valiosa fonte de informações para estudar o que ocorre no dia-a-dia. Muitas vezes os utilizadores utilizam o Twitter por razões pessoais, profissionais ou até com o intuito de relatar eventos da vida real. O objetivo desta tese é desenvolver um sistema capaz de prever o retorno das ações que compõem o DJIA sendo o principal indicador os tweets publicados por uma comunidade financeira. Numa primeira fase detetamos eventos especiais na vida das empresas por meio de uma análise ao sentimento implícito nos tweets extraídos. Numa fase seguinte implementamos um algoritmo genético com o objetivo de obter a melhor solução para o problema de otimização de portefólio. Ou seja, um algoritmo capaz de proporcionar a estratégia perfeita para selecionar um conjunto de ativos para obter o retorno esperado minimizando o risco. Numa fase final implementamos uma estratégia baseada na popularidade das empresas com o objetivo de obter os melhores retornos. Os resultados mostraram que os tweets recolhidos através da comunidade financeira definida permitem detetar eventos importantes na vida das empresas. Estes eventos detetados permitiram implementar um algoritmo genético que apresentou um retorno de 0,25%. Este retorno obtido apresentou um valor acima da média quando comparado com a aplicação da estratégia Buy&Hold ao DJIA que teve um retorno de -4,2% para o mesmo período. Por último, a análise da popularidade permitiu obter um retorno de 8,13% baseado no volume de tweets positivos de cada empresa.

Palavras-Chave: Preço da Ação, Índice Dow Jones, Twitter, Análise Sentimento, Deteção de Eventos, Algoritmo Genético

Abstract

The growing number of Twitter users makes it a valuable source of information to study what is happening right now. Many times users make use of Twitter for personal or professional reasons or even in order to report real life events. The objective of this thesis is to develop a system capable of predicting stock returns of the DJIA based on tweets posted by a financial community. In a first stage we detect special events in the life of companies through an analysis of the implicit sentiment in tweets extracted. As the next step we implement a genetic algorithm in order to get the best solution for the portfolio optimization problem. That is, an algorithm able to provide with the perfect strategy to select a set of stocks that give the expected return while minimizing risk. In the final phase a strategy is implemented based on popularity of each company in order to get the best returns. The results showed that the tweets collected by the financial community allow us to detect important events in the life of companies. These detected events were used to implement a genetic algorithm that had a return of 0.25%. This return obtained showed an above average value when compared to the application of the Buy&Hold strategy to DJIA that had a return of –4.2% for the same period. Finally the analysis of the popularity enabled us to get an 8.13% return based only on the volume of positive tweets for each company.

Keywords: Stock Price, Dow Jones Industrial Average (DIJA), Twitter, Sentiment Analysis, Event Detection, Genetic Algorithm

Table of Contents

List of Tables

List of Figures

List of Abbreviations

List of abbreviations (sorted alphabetically):

AI - Artificial Intelligence
ANEW - Affective Norms for English Words
ANN - Artificial Neural Networks
API - Application Programming Interface
B&H - Buy and Hold Strategy
CEO - Chief Executive Officer
DAX 30 - Deutsche Aktien Xchange 30
DJIA - Dow Jones Industrial Average
DMA - Derivatives Moving Average
EMA - Exponential Moving Average
EMH - Efficient Market Hypothesis
FTSE 100 - Financial Times Stock Exchange 100 Index
GA - Genetic Algorithm
GPOMS - Google-Profile of Mood States
HMA - Hull Moving Average
HTTP - Hypertext Transfer Protocol
JGAP - Java Genetic Algorithms Package
MACD - Moving Average Convergence Divergence
MaxEnt - Maximum Entropy Modeling
OBV - On Balance Volume
REST - Representational State Transfer
ROC - Rate of Change
ROI - Return on investment
RSI - Relative Strength Index
S&P500 - Standard&Poor's 500 Index
SMA - Simple Moving Average
SOFNN - Self Organizing Fuzzy Neural Networks
SVR - Support vector machine
TSI - True Strength Index

Chapter 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. Stock markets represent to investors a way to get profitability although any investment has an associated risk which makes it impossible to guarantee a certain profit. The 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 or social networks. Studies by (Hirshleifer, 2001), (Chan, 2003), (Vega, 2006) and (Tetlock, 2007) have shown that both the informational and 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 are 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 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 capable of predicting stock returns of the DJIA based on tweets posted by a financial community. First the algorithm detects the special events in the financial community based in the information from twitter. Based on events detected a genetic algorithm is implemented in order to predict the evolution of the Dow Jones index analyzing the possibility of the popularity of a company on Twitter be an indicator able to predict the market and get positive returns.

The methodology is composed of several stages, from the construction of the financial community Twitter, the collection of tweets published by the financial community, the analysis of the content of tweets and the detection of special events. In this methodology, 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. In the last stage of this thesis, the objective is to implement a genetic algorithm in order to present optimal solutions to the problem of predicting stock returns.

1.1 Motivation

Online social networks give the opportunity to hundreds of millions of users around the world to produce and consume content. In addition to online social networks provide access to a vast source of information still has an important role in the dissemination of information by increasing the spread of new information and points of view. Twitter is one of the most popular social networks in the world by with a large number of users. Through the publication of tweets users can share a variety of information and still express opinions about products, organizations or even people.

The motivation of this work results from the vast popularity and credibility that the social network Twitter has been having over time. Building a financial community in Twitter by which it is possible to extract tweets containing information that influence the market trend is one of the motivating challenges of this thesis. Sometimes the restriction of 140 characters to tweet require the user, makes them to use emoticons, URLs, hashtags, abbreviations, among others to be able to express in the best way but causing noise. So, as our database is composed of tweets, handle large amounts of data and high levels of noise is a challenging task for this project.

With the increasing of Twitter, it is normal that investors use this social network to make personal outpourings and publish opinions relating to the financial market. The investment portfolio optimization problem is the selection of a set of assets that give the investor an expected return while minimizing risk. Thus, the investment portfolio may be composed of various assets of the financial market, in different amounts. The implementation of strategies to be possible to find the best solutions for this portfolio optimization problem is a big challenge that many investors try to overcome over the years. In this thesis we face this challenge and to attempts to exceed we using genetic algorithms to select the assets of the Dow Jones index that allow us to obtain the best returns.

1.2 Contribution of this Master Thesis

This project presents an original work with respect to event detection and also the implementation of an evolutionary algorithm. There are six main contributions.

- ✓ The algorithm that defines the financial community inside Twitter and extracts all tweets for each user.
- ✓ 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.
- ✓ 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.
- ✓ Manual validation of events that have been characterized by large positive and negative peaks and are directly related to special events that positively and negatively affected the financial community.
- ✓ Investment strategy to predict the trend of the Dow Jones index through implicit sentiment in tweets and implementation of a genetic algorithm.
- ✓ Creating a strategy based on the analysis of the popularity of each company. This popularity is analyzed by the volume of positive tweets over the year. Based on the popularity of year, we chose companies to invest in the next year.

1.3 Organization of the Thesis

In this section we give a brief overview of the rest of this thesis.

Chapter 2 addresses the theory behind the developed work, namely some concepts relating to market analysis as technical analysis and fundamental analysis. Some works carried out by different authors in the field of evolutionary algorithms are presented in this section. 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.

Chapter 3 illustrates the system architecture and respective function of each module. In this section, for each module explained, we describe in detail what was done, what was implemented and tools used for the implementation of the system. Initially we discuss the social network Twitter, the basic operation and also some applications for developers. We explained in detail the phase of extracting data by implementing a financial community, data filtering stage, the sentiment analysis tools implemented for calculate the scores of tweets, the normalization realized, event detection and lastly the implementation of the genetic algorithm.

Chapter 4 proposes the validation procedure used to evaluate the developed strategy, where present three case studies that prove the event detection by the proposed methodology. The fourth test conducted presents a case study that shows the evolution of the genetic algorithm and the optimal solutions found in order to achieve positive returns for the thirty companies in the Dow Jones index. Finally we present a case study in the context of predicting stock returns based on the popularity of companies.

Chapter 5 summarizes the provided document, supplies the respective conclusion and the proposal of possible future work.

Chapter 2 Related work

This chapter presents background information and the literature review that is relevant to the development of this project. In the first part of the chapter (section 2.1) is presented a brief description of the theories that analyze the market. In section 2.2 is presented the literature review about the use of genetic algorithms to predict accurately the stock index. In section 2.3 several studies are presented about the Sentiment Analysis. In this section we discuss the tools used by the authors in order to be able to analyze the sentiment often implicit in the texts, publications or news. In Section 2.4 we give a general approach to social network Twitter, their influence on the financial market and also some literature that uses the social network for studies that analyze the sentiment of tweets. Section 2.5 is about event detection and explores several works that use the social network Twitter to try to detect special events. Finally the last section of this chapter, section 2.6, presents a table with a summary of the important literature presented along the chapter.

2.1 Market Analysis

The concept of efficient markets hypothesis (EMH) 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.

According to EMH stock prices follow a random walk stating that it is impossible to be able to predict if selling a declining investment before the end of the holding period is a better choice then to wait until the end of the holding period as in the buy-and-hold strategy. By selling before the end of a holding period the investor protects him/herself from further losses, but also deprives him/herself the potential stock price improvement during the remaining time of the holding period. Supporters of the EMH still claim that buy-and hold is superior to active portfolio management strategies (Malkiel, 2003). They dismiss active portfolio management strategies and as a result even stop-loss rules as pointless, inefficient and even wasteful. Instead they advise investors to stick to the buy-and-hold portfolio strategy. The Buy and Hold (B&H) portfolio strategy became widely acknowledged after the publication

of (Fama, 1970) where his study on the efficiency of the capital markets concludes that the B&H strategy was superior to active portfolio management in terms of return, risk and transaction cost.

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 or technical analysis. These tools are used by professional or amateur speculators to analyze the movement of prices of some financial assets. The main factor influencing the stock price is supply and demand.

Over time, there has been much research interest in predicting the index of stock prices. In addition to the traditional analysis, many studies have been performed using data mining techniques. In the data mining process (Thakkar, 2007), analysis and extraction of knowledge is carried out seeking to consistent standards and/or systematic relationships between instances of these data. To implement this technique, automated methods based on artificial intelligence, such as neural networks (ANN), genetic algorithms (GAs), among others, are used in order to improve the analysis process. In the next section we present some literature in the area of genetic algorithms for giving evidence that these algorithms can be used when the objective is to predict the index of stock prices.

2.2 Genetic Algorithm

Many studies on stock market prediction using artificial intelligence (AI) techniques were performed during the past decade. These studies used various types of Genetic Algorithms (GAs) to predict accurately the stock index and the direction of its change.

The main motivation for the use of genetic algorithms for classification rules discovery is also on better interaction between attributes provided by the GAs compared to algorithms based on ambitious strategic for classification rule induction, and that are generally used for data mining (Freitas, 2003). Another important advantage of using GAs for this task is the global search performed by this technique, increasing the likelihood of obtaining a set of rules with high predictive accuracy.

GAs are computational methods to simulate the evolutionary behavior of the species based on a Darwinian theory, whose goal is to optimize a given fitness function. These algorithms have been designed by (Holland, 1975), with the initial objective to study phenomena related to the adaptation of species and natural selection that occurs in nature (Darwin, 1859), as well as develop a way to incorporate these concepts to the computers (Mitchell, 1997). The GAs have a wide application in many scientific areas, including optimization problems can be mentioned solutions, machine learning, development strategies and mathematical formulas, analysis of economic models, engineering problems, various applications in biology as simulation bacteria, immune systems, ecosystems, format discovery and properties of organic molecules. Later, in 1989, (Goldberg, 1989), a student of Holland got his first success in industrial applications with GAs. Since then the GAs are used to solve optimization problems and machine learning.

Many previous studies have proposed many hybrid models of ANN and GA for the method of training the network, feature subset selection, and topology optimization. In most of these studies, however, GA is only used to improve the learning algorithm itself. Another approach to genetic algorithms involves feature discretization and the determination of connection weights for artificial neural networks (ANNs) to predict the stock price index. (Kim & Han, 2000) employed the GA not only to improve the learning algorithm, but also to reduce the complexity in feature space. GA optimizes simultaneously the connection weights between layers and the thresholds for feature discretization. Experimental results show that GA approach to the feature discretization model outperforms the other two conventional models.

The GAs are also subject of study when the objective is to implement new investment strategies for the stock market. (Simões, Neves, & Horta, 2010), proposed a new investment strategy focusing on the S&P500, FTSE 100, DAX 30 and the Nikkei 225 in the medium/long term (2004 -2009). This new strategy uses genetic algorithms to calculate the perfect strategy using the combination of two technical indicators (Simple moving average (SMA) and Derivatives moving average (DMA)). Another article developed in the area was created by the authors (Gorgulho, Neves, & Horta, 2011) and presents a new also based approach in intelligent computing, in particular GAs, which aims to manage a financial portfolio using technical analysis indicators (EMA, HMA, ROC, RSI, MACD, TSI, OBV). The results were tested regarding the return on investment (ROI) achieved by the strategies considered in

the years 2003 to 2009. The approach allowed to prove the superiority of the system using the GAs based on technical signals.

The use of genetic algorithms, neural networks and genetic programming in an attempt to find a low cost solution is very common. To predict accurately the stock index, we have to really understand what influences the financial market. The main factor influencing the stock price is supply and demand. But there are other aspects that influence this value. In the information age, the news can spread throughout the world, sometimes at a higher speed than happens. However, the news is not the only means of disseminating information that inflame the financial market. Reviews and publications that daily invade social networks also allow us to extract very specific information about the kind of motivations of the people who produce them. These motivations may have implicit positive or negative feelings. For all the reasons presented here it is important to give importance to an increasingly exploited technique, Sentiment Analysis. The following section presents some literature around this technique has been increasingly used to realize the feeling that people often transposed into the content produced.

2.3 Sentiment Analysis

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. The opinions have great influence on people's behavior. Simple decisions taken by the people as what the film will see which product to choose or even what stocks are best to invest are based often on opinions given by others. This is not only true for people, but also for organizations that have seen the notion of customer feedback about their products and services as an added value to organizations. Organizations base their business strategies and investments in the opinion of customers about your products or services. The importance of the opinion is so great that many companies (e.g. marketing, public relations, and research) have their business aimed at obtaining this type of information. Traditionally, the answer to questions involving the public involves techniques such as field research, phone calls or written questionnaires. These techniques involve costs, are restricted to a well-defined group or sample, and your return is time consuming and often ineffective. The explosion of social networks has changed this scenario, providing individuals and organizations contents of diverse opinion and in large volumes. Web users have the opportunity to record and disseminate their ideas and opinions through comments, discussion forums, blogs, Twitter, social networks, among others. This increases the options of individuals in search of opinions, as they are no longer limited to your personal network of contacts (e.g. family, friends and professional connections) or opinions publicly available specialists (e.g. magazines, newspapers). However, the large volume of

information produced daily implies a need for methods and able to automatically process not only tools the content of publications, but also the opinion and sentiment they express. Mining opinions or sentiment analysis is the computational study of opinions, evaluations, attitudes, affections, visions, feelings and emotions expressed in the text, (Liu, 2012) (Pang & Lee, 2008).

Several subtasks can be identified within sentiment analysis, (Tsytsarau & Palpanas, 2012):
1) Determine the subjectivity of the document: often called subjectivity rating. This subtask determines whether a text is objective (expressing a fact) or subjective (expressing an opinion or emotion).
2) Determine the orientation of the document: often called sentiment classification. This subtask determines the polarity of a text. In other words, it determines if the text expresses a positive feeling or a negative feeling about your subject.
3) Determine the strength of the orientation of the document. This subtask decides whether the positive feeling expressed by a text is weakly positive, moderately positive or strongly positive.

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, 2013) 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.

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.

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.

2.4 Twitter for Market Analysis

Twitter is an online social media service used by millions of individuals and organizations worldwide to exchange short messages of up to 140 characters (tweets). It has rapidly evolved over the past few years to become a complete ecosystem and a powerful tool in several areas such as news, politics, health, and in our case, finance. Each tweet contains a text message with additional embedded metadata such as author, time and date, location, and language. Today, Twitter is a great source of information. 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.

An example that demonstrates the growing popularity of Twitter and its credibility as a social network was when Wall Street showed interest in its platform. On the 4th of April 2013, (Bloomberg) announced that it is the first financial information platform that integrates real-time Twitter feeds directly into the investment workflows of market professionals. The head of sales and product development for the Bloomberg Professional service, Jean-Paul Zammitt, explains in a statement "When important news is shared on Twitter, traders and investors need to be able to access it, and validate its importance in order to incorporate that information into their decision-making process."
With the fast growth of Twitter popularity, it is important to note that not only individual users have interest in this platform but also organizations, businesses and public services. The authors, (Kaplan & Haenlein, 2010) discuss the reasons for the increasing success of social networks. The study concludes that Twitter's success comes from its unique communicational characteristics. These communicational characteristics that allow the spread of information on several subjects in real-time, enable the user/consumer quick access to positive or negative information about a particular product or service. An event that demonstrates the rapid spread of content produced on Twitter and its influence on the financial market, took place on 23 April 2013. This event proved that Twitter not only influences the organizations and individual companies as well as the market in general. On this day, a group of hackers assumed control over Associates Press' Twitter account and posted the following message: "Breaking: Two Explosions in the White House and Barack Obama is injured". This triggered a 0.9 percent immediate decline in the S&P 500, wiping out about $136 billion in market value from the companies in the index. The market recovered within three minutes as investors determined that the tweet was false. Some traders said that the big drop derived from the publication of the tweet may have been caused by algorithm trading robots tracking the news headlines, reacting contrary to humans, who would have realized that the information was false before trading on it, (Kisling, Lam, & Mehta, 2013).

Because of such incidents in order to be less likely to happen, one way to overcome this challenge is to require verification of a second unique and reliable source before trading. These algorithms used in the market should take special care when implemented and must be accompanied of powerful sentiment analysis tools in order to minimize such incidents. Due to Twitter reflects a great influence on the market and the influx of users is increasingly high, numerous studies have been conducted in the context of analyzing the implicit sentiment in the content published on Twitter. (Bollen, Maoa, & Zengb, 2011), investigated whether measurements of collective mood states derived from large-scale

Twitter feeds are correlated to the value of the 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). Following the publication, this paper launched much of the current studies on the relationship between Twitter and market sentiments. Another similar paper (Mittal, Anshul; Goel, Arpit, 2012), they applied sentiment analysis and machine learning principles to find the correlation between “public sentiment” and “market sentiment”. They used twitter data to predict public mood and used the predicted mood and previous days’ DJIA values to predict the stock market movements. The authors proposed a new cross validation method for financial data and obtain 75.56% accuracy using Self Organizing Fuzzy Neural Networks (SOFNN) on the Twitter feeds and DJIA values from the period June 2009 to December 2009. Their work is similar to (Bollen, Maoa, & Zengb, 2011), with a few minor modifications.

The intentions of Twitter users are different. Some people use Twitter only as a means of conversation, to talk about their daily activities, professional use, and participation of productive information or sometimes malicious share information. To understand the influence of users on Twitter, (Yang, Steve Y.; Mo, Sheung Yin Kevin; Zhu, Xiaodi, 2013) 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 by Yang et al. could be concluded that the sentiment generated by each node of the financial community has predictive power in consistent market returns and own 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) present a new approach for automatically classifying 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.

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.

2.5 Event Detection

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.

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.

(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.

(Gomide, Lima, Gomide, Roque, & Silva, 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). Finally, (Petrović, Osborne, & Lavrenko, 2010) used locality-sensitive hashing to detect the first tweet associated with an event in a stream of Twitter messages.

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.

Next is shown the last section of this chapter which briefly details the most relevant literature for the realization of this work.

2.6 Overview approaches to work purposed

To conclude the background work chapter, a resume table is shown with different pattern recognition approaches documented on recent papers and their performance.

Table 1 - Recent pattern recognition papers and their performance.

Chapter	Reference	Date	Period of Simulation	Market Tested	Algorithms	Tool for analyze sentiment	Best Results
Literature of Genetic Algorithms	(Gorgulho, Neves, & Horta)	2011	6 Jan. 2003 to 6 Jan. 2009	DJIA	GA	-	ROI(%) → 62.95
Literature of Sentiment Analysis	(Schumaker & Chen)	2009	During 5 weeks	S&P 500	SVM	AZFinText system design	Return (%) = 2.06
	(Nielsen)	2011	-	-	SentiStrength	ANEW Vs New Word List	New word list perform better than ANEW
	(Geva & Zahavi)	2013	During 8,5 months	S&P 500	ANN	Digital Trowel	Return (%) = 6

Chapter	Reference	Date	Period of Simulation	Market Tested	Algorithms	Tool for analyze sentiment	Best Results
Literature of Twitter	(Go, Bhayani, & Huang)	2009	6 April 2009 to 25 June 2009	-	MaxEnt	-	Accuracy(%) = 83
	(Bollen, Maoa, & Zengb)	2011	28 Feb. 2008 to 19 Dec. 2008	DJIA	SOFNN	OpinionFinder Google-Profile of Mood States	MAPE (%) = 1.83 Direction Accuracy (%) = 86.7
	(Mittal, Anshul; Goel, Arpit)	2012	June 2009 to Dec. 2009	DJIA	SOFNN	-	MAPE (%) = 11.03 Direction Accuracy (%) = 75.56
	(Yang, Steve Y.; Mo, Sheung Yin Kevin; Zhu, Xiaodi)	2013	15 Oct. 2013 to 15 Nov. 2013	DJIA	Algorithm developed by authors	Dictionary of (Hill & Liu, 2004)	Using critical node with betweenness centrality has more impact.
Literature of Event Detection	(Mendoza, Poblete, & Castillo)	2010	27 Feb 2010 To 2 March 2010	Chile	Algorithm developed with Hashtag's	-	Veracity of tweets (%) = 95.5
	(Wang, Can, Kazemzadeh, Bar, & Narayanan)	2012	12 Oct. 2011 to 29 Feb. 2012	DJIA	Naive Bayes	-	Avg. Accuracy (%) = 59
	(Gomide, Lima, Gomide, Roque, & Silva)	2014	1 Dec 2010 To 31 May 2011	Brazil	Alpha de Cronbach/ Spearman Correlation and Cluster Analysis	-	Spearman correlation = 0.924 (high positive correlation)

Chapter 3 Methodology

The proposed system presents a new approach to predict the trend of the Dow Jones index based on sentiment analysis of the tweets posted by a financial community.

First, Section 3.1 shows the general architecture of the proposed system as well as a brief explanation of each module. The following subsections provide a detailed explanation of each module, the choices made for the elaboration as well as the tools used and some implementation decisions.
Sub Section 3.1.1 gives an insight into the Twitter universe and the functions available to the user.
In this subsection 3.1.2 we discussed the important task of this system, the data extraction. Basically in this subsection we define a financial community and explain all community building steps. The database of this project is the tweets produced by the financial community created.
Extracted the tweets of the financial community, in subsection 3.1.3 present the filters implemented for removing noise in tweets in order to get only the important tweets for this work.
Subsection 3.1.4 presents the four sentiment analysis tools implemented to obtain a positive, negative or neutral score for the tweets.
Subsection 3.1.5 it is explained the implementation of normalization module that is responsible for normalizing the volume of tweets over time in order to provide consistency in total of tweets extracted.
Subsection 3.1.6 is set the choice made for the detection of events in companies over time based on the sentiment evolution.
Finally in this chapter is presented to subsection 3.1.7 which provides a brief introduction to genetic algorithms and the definition of genetic operators that we will use in this proposed system.

3.1 System Architecture

The proposed system uses an investment model bringing as an indicator the views of users published on Twitter about the financial market, with particular emphasis on companies that make up the Dow Jones index. The objective of this system is to analyze the content produced by a financial community on twitter social network, detect possible events over time in the lives of companies and through these events detected choose the stocks that should be included in the portfolio with the ultimate goal of get positive returns.

The system architecture is shown in Figure 1 and is composed of several main modules whose functions are detailed in the following sections.

Initially it is defined a financial community chosen by us from the twitter universe that will be used to extract the 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. With tweets organized into files where each file is one of thirty companies, our system will use the sentiment analysis block to calculate the daily score for each tweet. This block consists of four sentiment analysis tools that will evaluate the tweets individually. In an intermediate stage of the system, is

necessary make a data normalization in order to standardize the volume of tweets. As the flow of tweets has increased over time largely due to the increased number of users, the normalization module is indispensable in this system.

At this stage, we have all tweets organized into files (30 files), with the respective score for each sentiment analysis tool implemented and with their values normalized. With all this information, we can analyze the daily sentiment of users about the company over time. With the Event Detection module we will analyze the daily scores of companies and through very positive or very negative days we will try to see if something good or bad happened in the company. The module that detects events as well as stock quotes are accessed by Genetic Algorithm module. In this module, the user also provides information such as the initial budget, the test period and the strategy that the algorithm will perform. GA is the module that starts executing and generates the initial population of individuals (chromosomes). Each individual is submitted to a simulation using the Investor Simulator module for a quality check. During the training period there is an evolution process. Based on the evaluation, it selects the chromosomes to reproduce and apply the methods of crossover and mutation to create new individuals. The survivals, this is, the individuals presenting higher returns on fitness function, are selected to constitute the new generations. Once the training period is over, the Individuals are sorted by performance. Evaluated the performance of the population and when the training ends, the top five individuals are selected for a real simulation. With the best five chromosomes, the system reruns the Investor Simulation module this time with the parameterized period for the real test (in this case is the following year). After testing, the results are delivered to the user via the Results module.

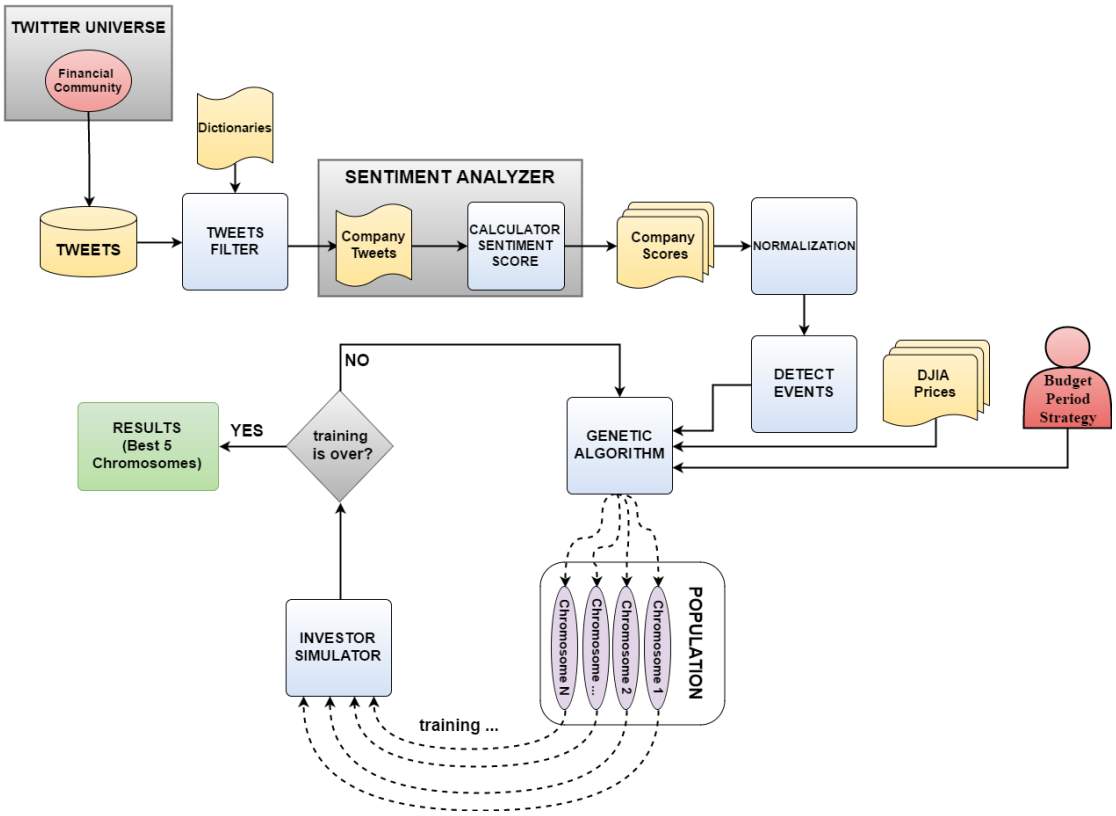

Figure 1 – System Architecture.

3.1.1 Twitter Social Network

Since its official launch in July 2006 to the present day, Twitter has become one of the most important social networks present on the web. With hundreds of millions of users, Twitter is a service that allows users to write and share quick text messages limited to 140 characters, known as tweets. These messages can differ greatly, approaching from more personal topics or until news or reactions to important social and international events. To receive tweets from other users in your personal area, a user must "follow" others, a concept that differs from the friendly relations of other social networks. For example Facebook is a social network where the user needs to ask to be friends and the friend needs to accept. The Twitter concept is different because a user can "follow" other without the reverse is the case. 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 user profile customizations, such as changing the background image, colors, and fill in personal data since each user have a web page with your individual area.

In twitter every user has a unique screen name and a unique numerical ID. In addition to this personal information, it is still possible to know a lot about the user, such as (Figure 2):

- ✓ Name
- ✓ Screen Name
- ✓ Follower count
- ✓ Following Count
- ✓ Follower List
- ✓ Following List
- ✓ All tweets from a particular user
- ✓ Location

Figure 2 - Example of Twitter User.

The tweets posted by users can be grouped by *hashtag* words preceded by the # character, *at* preceded by the @ character and *cashtags* preceded by the $ character, see Table 2. In addition, users can make retweet in a tweet posted by other users. The Retweet happens usually when a

person writes a sentence about something that is interesting to others, or when it is a matter of public interest, which must be passed forward.

Table 2 – Twitter terms and concepts.

TWITTER TERMS AND CONCEPTS	
Tweet	An up to 140 character long text message
Hashtag (#)	Identifies a topic (e.g. #earnings)
At (@)	Identifies a username (e.g. @CNBC)
Cashtag ($)	Identifies a stock ticker (e.g. $AAPL)
Retweet	When another user relays your tweet to their followers

The information we can extract a tweet can be:

- ✓ Tweet content
- ✓ Tweet id
- ✓ Tweeter id
- ✓ Location
- ✓ Time
- ✓ Mentions
- ✓ Trends
- ✓ Hashtags
- ✓ Cashtag
- ✓ Number of retweets

Twitter was the social network chosen to serve as a database for this thesis. All information at the user level as the tweet level can be extracted with the help of various applications that Twitter provides to developers.

➢ *Twitter API for Developers*

Like all Web Applications, social networking sites are hosted on web servers. For other applications to be able to interact with Web Applications, are used Web Services. So Online social networks APIs are Web Services designed to offer features and functionalities, present in the social network, to external applications. These APIs transform the social networking sites on platforms for social applications. The Online Social Networking Web APIs are web services that allow access and manipulation of social information. Most APIs are implemented through the REST (Representational State Transfer)

principles. In this technique the exchanged messages are encapsulated directly in the Hypertext Transfer Protocol **(**HTTP).

The Twitter API allows multiple applications to connect to it for various purposes. The Twitter API can be divided in three parts:

- (REST APIs): This is the main API. It is the one that has more services and is intended for handling user data and connections between them, as well as sending messages. This API provides programmatic access to read and write Twitter data. The REST API identifies Twitter applications and users using authentication: OAuth.
- (Search API): The Search API provides services to search for messages and users. Returns a collection of relevant Tweets matching the specified query. Not all Tweets will be indexed or made available via the search interface.
- (Streaming APIs): This API has an architecture a bit different from the first two. It aims to generate persistent connections to exchange information synchronously. This is useful for desktop systems, or in systems that need constant updates to the message history. If the intention is to conduct singular searches, read user profile information, or post Tweets, consider using the REST APIs instead.

Many services present in the API require authentication and authorization by the user. Twitter uses the open authentication standard, OAuth. It is an authentication protocol that allows users to approve the application to act on their behalf without sharing their password.

Initially you need to register in the application. This is because before starting the flow of communication, it is necessary to generate an API Key, a Consumer Key and Consumer Secret. After the authentication is properly performed, is already possible to start making use of Twitter application, once Twitter can already identify our application.

➢ *Twitter4J Library*

As already said, nowadays with the large number of data available on social networks is very useful to extract information and data in an agile way. In this context, the development of systems that make it possible to obtain these data, correlate them and turn them into useful information is of great value for many lines of business. Faced with this challenge, this thesis proposes the use of Twitter4J library to enable integration with Twitter and so explore its many features aimed at extraction, search and data analysis as well as automated forms of interaction increasingly effective. Below are some simple methods that this library provides (the methods highlighted in bold are the used in this database extraction):

For users:

- ✓ **The list of his followers;**
- ✓ **The list of users they follow;**

- ✓ The total number of published tweets;
- ✓ **The last 3200 tweets.**

For Tweets:

- ✓ **The actual message;**
- ✓ **The publication date of the tweet;**
- ✓ The number of retweets that has so far;
- ✓ If the tweet itself is retweet.

Twitter4j is an unofficial Java library for the Twitter API. It is an open source library created by (Yamamoto) that easily helps integrate our Java application with the Twitter service. This library has methods which allow us to use all the three API of the Twitter API. In this project, for the extraction of data, we will use the Twitter REST API v1.1 in parallel with the library twitter4j. The use of the Twitter REST API is limited, so their applications cannot be connected to any number of times to ask anything, this situation is called Rate Limit. In this project we had the limitation of having to wait 15 minutes every 180 gets. This limitation delayed our data collection and made it impossible for us to have a database higher due to the proposed deadline for obtaining the data.

3.1.2 Data Extraction

The first and most important step for the extraction of data was choosing which Twitter users to take into consideration to create our financial community. The entire community selection process is detailed below.

➢ *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 users in Twitter universe that somehow are related to the financial market.

Construction of the Financial Community

The financial community proposed in this work begins with a collection of eleven user accounts representing investment experts, financial news providers, managers and founders of companies. The selection criteria for these eleven users is based into persons that are very known in the

financial community. The second stage of the selection of user accounts was based on the followers of this community of eleven 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. In this stage only users who have more followers than users that they follow were chosen. In **Figure 3** is shown an illustration of this selection of 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 and will be part of our database. Below is a list of the eleven chosen personalities and their functions to realize a bit more as it was elaborated the financial community.

1. **Jeffrey Gundlach** has been the Chief Executive Officer at DoubleLine Capital since December 2009. He was the founder of Doubleline Capital, an investment firm.

2. **Bill Ackman** is an American hedge-fund manager. He is the founder and CEO of Pershing Square Capital Management LP, a hedge-fund management company. He considers himself an activist investor.

3. **Jeremy Grantham** is a British investor and co-founder and chief investment strategist of Grantham, Mayo, & van Otterloo (GMO), a Boston-based asset management firm. GMO is one of the largest managers of such funds in the world, having more than US $118 billion in assets under management as of March 2015.

4. **John Sculley** is an American businessman, entrepreneur and investor in high-tech startups. Sculley was vice-president (1970–1977) and president of Pepsi-Cola (1977–1983), until he became chief executive officer of Apple Inc. on April 8, 1983, a position he held until leaving in 1993.

5. **Ray Dalio** is an American businessman and founder of the investment firm Bridgewater Associates. In 2012, Dalio appeared on the annual Time 100 list of the 100 most influential people in the world. In 2011 and 2012 he was listed by Bloomberg Markets as one of the50 Most Influential people. According to Forbes, he was the 30th richest person in America and the 69th richest person in the world.

6. **Ron Johnson** is the former chief executive officer of J. C. Penney. Previously, he was the senior vice president of retail operations at Apple Inc., where he pioneered the concept of the Apple Retail Stores and the Genius Bar, and the vice president of merchandising for Target Corporation, where he was credited for making the store "hip." He is currently the head of Enjoy, a startup company.

7. **Gary Shilling** is an American financial analyst and commentator who appears on a regular basis in publications such as Forbes magazine, The New York Times and The Wall Street Journal. He is President of A. Gary Shilling & Co., Inc., editor of A. Gary Shilling's Insight, and member of The Nihon Keizai Shimbun Board of Economists. He is featured frequently on business shows on radio and television, and as a recognized orator, addresses conventions of global business groups.

8. **Jeff Bezos** is an American technology entrepreneur and investor. He has played a role in the growth of e-commerce as the founder and CEO of Amazon.com, an online merchant of books and later of a wide variety of products and services, most recently video streaming. Amazon.com became the largest retailer on the World Wide Web and a model for Internet sales.

9. **Daniel Kottke** It is a computer scientist and one of the first employees of Apple Inc.

10. **Ronald Wayne** is a retired American electronics industry worker. He co-founded Apple Computer (now Apple Inc.) with Steve Wozniak and Steve Jobs, providing administrative oversight for the new venture.

11. **Adam Davidson** is an American journalist focusing on business and economics issues for National Public Radio. He is currently one of the co-hosts of the Planet Money podcast. Previously he has covered globalization issues, the Asian tsunami, and the war in Iraq, for which he won the Daniel Schorr Journalism Prize.

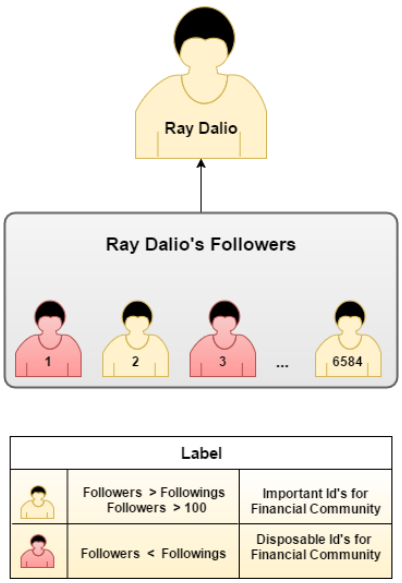

Figure 3 - Example of a Financial Community Subset.

With eleven Ids defined initially and with the help of Twitter4j library we developed a software for obtaining the financial community. Basically we use our personal account Twitter to follow the eleven personalities of our financial community. As the next step, through the functions of java library (Twitter4J) we obtained the users IDs of our followers. With the acquisition of the eleven ids was easy to get the rest of the financial community because twitter4j has functions that give us all the ids that follow the personality and even the followers of personality followers.

- ***Tweets Database***

With all the ids of the financial community collected, it is performed the last stage of data extraction. For each user, using the methods of Twitter4J library, we extract the last 3200 tweets published with the respective date of publication. At the end of this step we have a database composed of files. Each file is relative to a user and contains all the tweets published from September 2013 to September 2015.

3.1.3 Tweets Filter

At this stage we have a tweets file for each user of our financial community. The filtering is illustrated in Figure 4. With the database made, there was the need for a pre-processing of all content. The first filter makes use of a dictionary prepared by us containing hundreds of words related to the financial market and the stock exchange. 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).

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 the keywords of each company as can be seen in Table 3. 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 Table 3. 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.

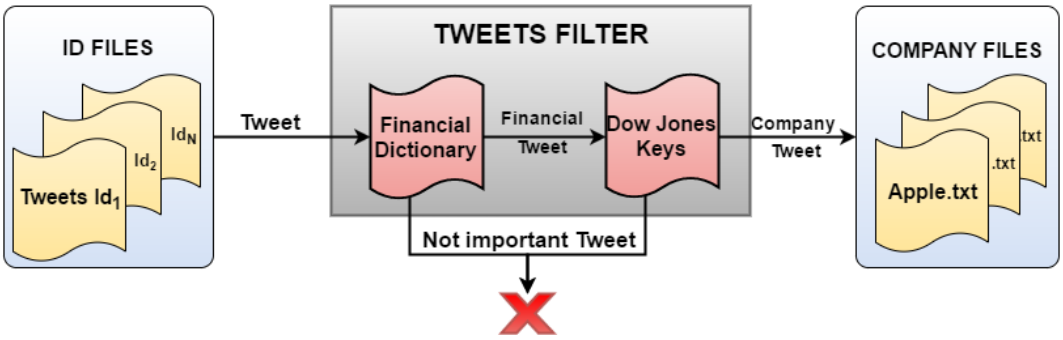

Figure 4 - Filtering Tweets.

Table 3 - List of keys DJIA that make up the second stage of the filter.

COMPANY	KEYWORDS		
Apple	APPLE	$AAPL	#APPLE
American Express	AMERICAN EXPRESS	$AXP	#AMERICAN EXPRESS
Boeing	BOEING	$BA	#BOEING
Caterpillar	CATERPILLAR	$CAT	#CATERPILLAR
Cisco	CISCO	$CSCO	#CISCO
Chevron	CHEVRON	$CVX	#CHEVRON
DuPont	DUPONT	$DD	#DUPONT
Walt Disney	DISNEY WALTDISNEY	$DIS	#DISNEY #WALTDISNEY
General Electric	GENERALELECTRIC GENERAL ELECTRIC	$GE	#GENERALELECTRIC
Goldman Sachs	GOLDMANSACHS GOLDMAN SACHS	$GS	#GOLDMANSACHS
Home Depot	HOMEDEPOT HOME DEPOT	$HD	#HOMEDEPOT
IBM	INTERNATIONAL BUSINESS MACHINES IBM	$IBM	#INTERNATIONALBUSINESSMACHINES
Intel	INTEL	$INTC	#INTEL
Johnson & Johnson	JOHNSON&JOHNSON JOHNSON & JOHNSON	$JNJ	#JOHNSON&JOHNSON
JPMorgan Chase	JPMORGANCHASE JPMORGAN CHASE	$JPM	#JPMORGANCHASE
Coca-Cola	COCA-COLA	$KO	#COCA-COLA

	COCA COLA		#COCACOLA
McDonald's	MCDONALD'S MCDONALDS MCDONALDS	$MCD	#MCDONALD'S #MCDONALDS
3M	3M MMM	$MMM	#3M
Merck	MERCK	$MRK	#MERCK
Microsoft	MICROSOFT	$MSFT	#MICROSOFT
Nike	NIKE	$NIKE	-
Pfizer	PFIZER	$PFE	#PFIZER
Procter & Gamble	PROCTER&GAMBLE PROCTER & GAMBLE	$PG	#PROCTER&GAMBLE
Travelers Companies	TRAVELERSCOMPANIES TRAVELERS COMPANIES	$TRV	#TRAVELERSCOMPANIES
United Health	UNITEDHEALTH UNITED HEALTH	$UNH	#UNITEDHEALTH
United Technologies	UNITEDTECHNOLOGIES UNITED TECHNOLOGIES	$UTX	#UNITEDTECHNOLOGIES
Visa	VISA	$V	#VISA
Verizon	VERIZON	$VZ	#VERIZON
Walmart	WALMART WAL-MART	$WMT	#WALMART #WAL-MART
Exxon Mobil	EXXONMOBIL EXXON MOBIL	$XOM	#EXXONMOBIL

3.1.4 Sentiment Analyzer

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 and the rest were developed by different authors. Next are described the tools implemented and in Table 4 is an example of the evaluation of six tweets to four different tools.

➢ *MySentiment API*

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. 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 only the respective words and punctuation. The bigram list consists of sets of two words and the respective score in this set. These dictionaries were created using a sample of 775.310 tweets between April and December 2012.

Our tool is developed based on the evaluation made in these dictionaries. Initially the sentence is read and the scores assigned to each word are added. In the next stage the phrase is read again but in sets of two words (bigrams). The final result is the sum of the scores of unigrams and bigrams found in the tweet.

➢ *TextBlob*

TextBlob 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, (Loria). 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) 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*

Affin 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 (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 4** an example of six tweets is presented with the appropriate calculation of scores for the four sentiment applications mentioned above. 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 4 - Example sentiment evaluation tweets.

Theoretical Analysis	Tweets	Score			
		MySentiment Api	Text Blob	Senti Strength	Affin
Positive	"This bigger pixels thing by Apple makes HTC look good. Very good."	**2,44**	**1,26**	**0,53**	**6**
Positive	"We live in an exciting period of transformation in #Technology - #MicrosoftHoloLens"	**3,32**	**0,22**	**0,64**	**3**
Negative	"#Apple stock is falling since today's announcement. Not surprised."	**-1**	**-0,05**	**-0,3125**	**-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"	**-12**	**-0,37**	**-0,15**	**-2**
Neutral	"Walmart informational meeting in GB over plans for downtown @CityofGreenBay @WFRVNews #walmart http://t.co/5ywrgJ2R72"	**0,078**	**0**	**0,437**	**0**
Neutral	"Catch me on CNN tomorrow at 6:40AM talking Apple's announcements. Tune In!"	**0,749**	**0**	**0,25**	**0**

3.1.5 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 **Figure 5** 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. **Figure 6** shows two graphs with an example for Apple Company to better understand the effect of normalization in the data. As shown in **Figure 6** 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.

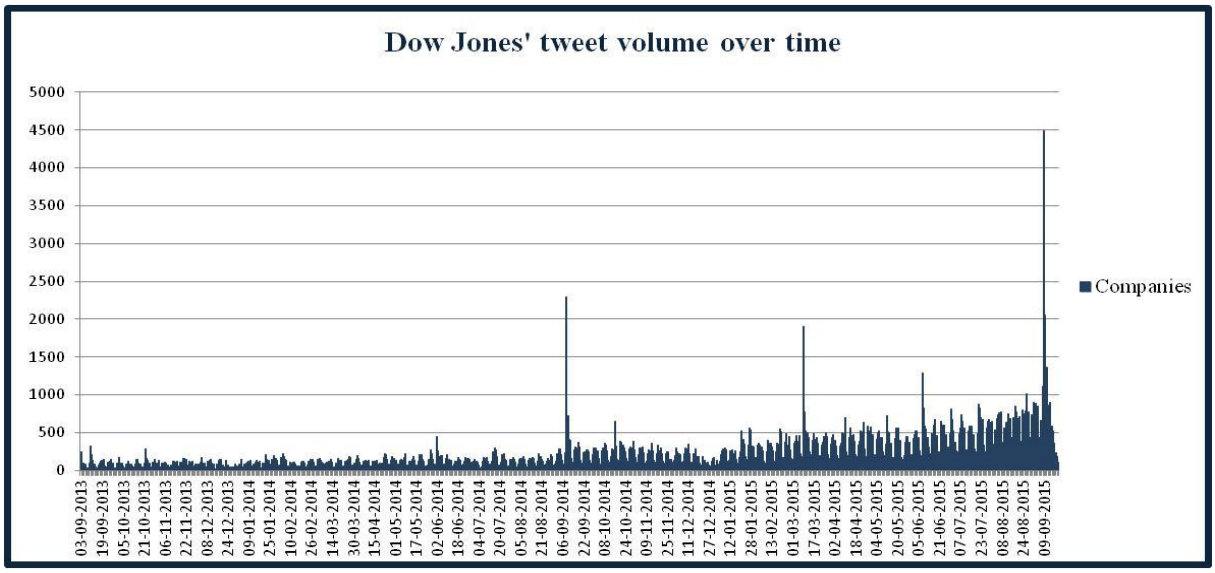

Figure 5 - Total tweets over time for DIJA.

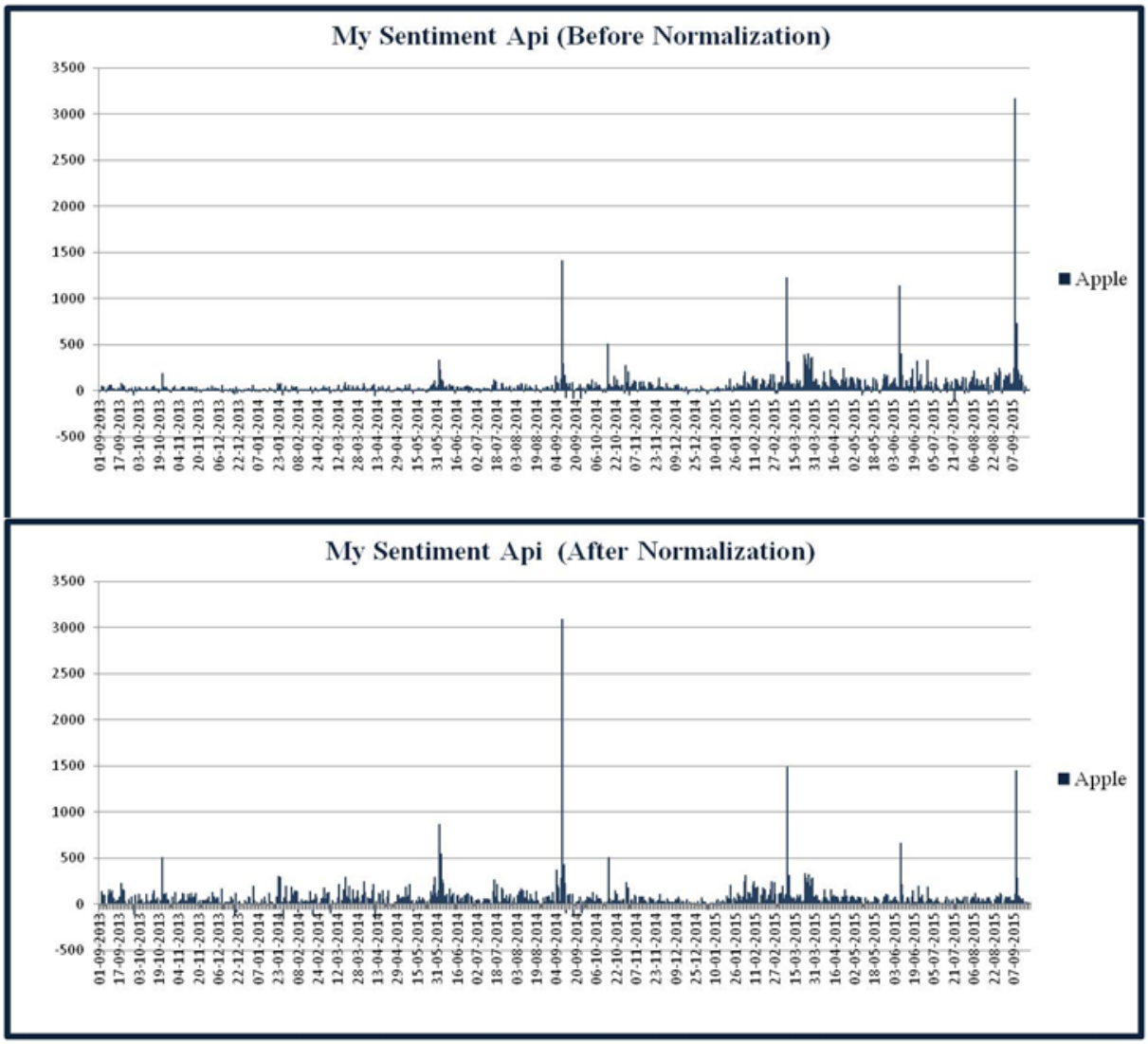

Figure 6 - Normalization of Apple's Company.

3.1.6 Event Detection

This module is responsible for detecting events during the company's test period. The event detection is based on analyzing the graph peaks over time. That is, throughout the evolution of the company for the test period, this module identifies the very negative or very positive peaks jumping to the graph view. These peaks correspond to a lot of positivity or negativity daily hence can correspond to an important event for the company. The module's responsibility is to identify whether actually happened a peak corresponding to a significant event for the company. This peak is the result of analysis to daily sentiment of tweets published concerning the company. The analysis of sentiment as stated earlier, is carried out with four different tools, hence resulting 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.

3.1.7 Genetic Algorithm

A genetic algorithm is a search technique used in computing to find exact or approximate solutions to search and optimization problems. Genetic algorithms are a particular class of evolutionary computation that uses techniques inspired by evolutionary biology. The big advantage of genetic algorithms is that you do not have to specify all the details of a problem in advance.

This thesis aims to use this type of algorithms characterized by its robustness, generality and easy adaptability in order to find optimal solutions to the proposed problem. The investment portfolio optimization problem is the selection of a set of assets that give the investor an expected return while minimizing risk. Thus, the investment portfolio may be composed of various assets of the financial market, in different amounts. For some investors, the mounting of an investment portfolio aims to decrease the risk associated with changes in the prices of assets. However, other investors, who prefer to gain a little higher, create their portfolio with more risk in order to obtain a higher level of return. There are also bold investors who assemble their portfolio targeting only the gains, ie, without giving the risk great importance. The investment portfolio optimization problem has a variable that only the investor can provide, which is the provision each has for risk. The use of genetic algorithms in order to seek a solution to a model of the stock portfolio optimization is useful for any type of investor, be it conservative, moderate or aggressive investor. In this work, we focus on the formulation of a new genetic algorithm with the simple goal of making a profit for the investor. The goal is to create an investment portfolio that presents us good returns based on the opinion of the users' publications about thirty companies that comprise the Dow Jones index. In this section is described the process of evolution and how genetic operators assist in the resolution of our problem.

➢ *Introduction*

Over time, the population evolves to ensure their survival. This evolution takes place in accordance with the principles of natural selection proposed by (Darwin, 1859). According to the theory proposed by Darwin, organisms of a population that are better suited to the environment in which they live are more likely to survive and reproduce than less adapted individuals that usually end up being eliminated. If there are some individuals with a combination of good characteristics is very expected that the descendants are further endowed with good characteristics. As already said in the section of the state of the art, these theoretical foundations in respect of genetic algorithms were developed by (Holland, 1975). **Figure 7** shows the structure of operation of a genetic algorithm.

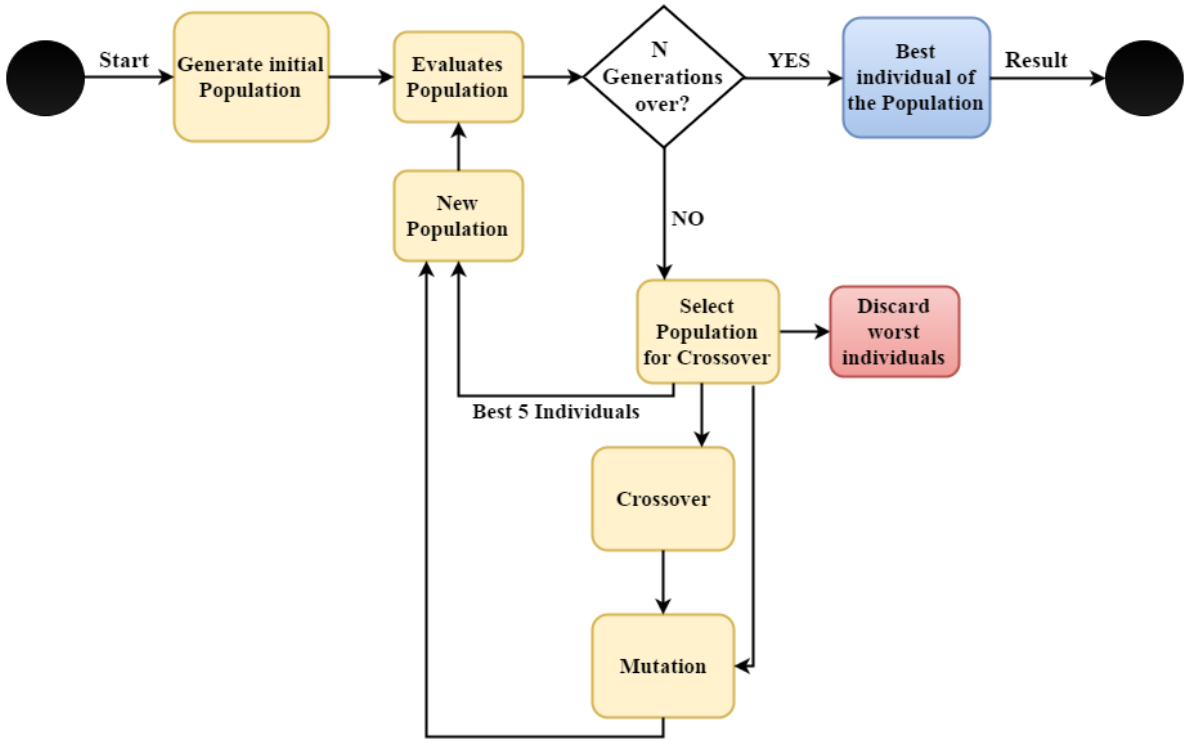

Figure 7 **-** Structure of operation of a traditional GA.

Basically a genetic algorithm simulates the process of natural evolution from a population of possible answers to the problem and then subjecting the population to the evolutionary process that comprises the steps of evaluation, selection, crossover, mutation, update and finally the finish.

The evaluation process is characterized by fitness evaluation of individuals of the initial proposal population. The next step is followed by the selection of individuals for reproduction process. The probability of a given solution is selected is proportional to its fitness. In the crossover phase characteristics of the chosen solutions are recombined thus generating new individuals. Mutation is the stage in which the characteristics of the resulting individuals of the reproduction process are changed, thereby adding variety to the population. The update phase is characterized by the integration of individuals created in the population. Finally the last finishing step verifies that the evolution of the closing conditions have been met, returning to the evaluation stage if this conditions have not been met.

To implement the genetic algorithm in this project we used the Java programming language because it has various libraries that allow the implementation of this family of algorithms. The selected library was JGAP (Java Genetic Algorithms Package) one of the most complete and largest community activity. In this project we use this JGAP library because it has the advantage of continuing to be updated and it is the library that has more maturity among many others for the same purpose. After presenting a definition and explaining in general the operation of genetic algorithms, it is necessary to discuss about the various processes running within it. So, the next topic is a study about such processes, called genetic operators.

➢ *Chromosome Representation*

A genetic algorithm starts from the execution of the operator named Initialization. This operator is the creation of an initial population which is basically the set of individuals that are being evaluated as a solution which will be used to create a new set of subjects for analysis. The size of the population can affect the overall performance and efficiency of genetic algorithms. Very small populations have great chances of losing the diversity required to create a good solution because they provide a little cover in search of the solution. On the other hand, if the population was made up of many individuals, the algorithm may lose much of its effectiveness for the delay to assess the fitness function at each iteration and may need to work with more computational resources.

The population is the group of multiple individuals or multiple chromosomes. An individual or chromosome is composed of a set of genes that are basically different variables that characterize the individual. **Figure 8** shows the chromosome model defined in this methodology. Our model chromosome consists of 11 genes, ie 11 different variables which will explain each of them next.

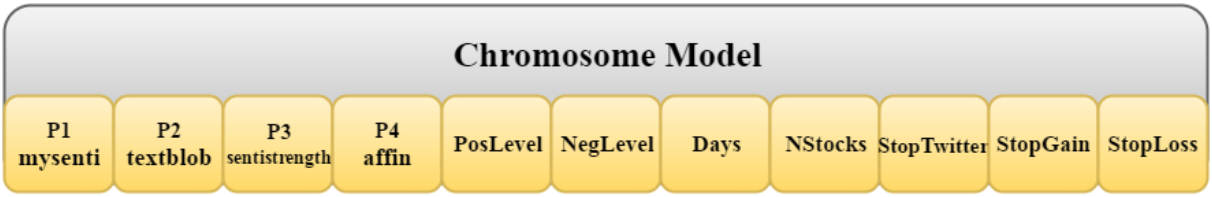

Figure 8 - Definition of the chromosome model used to implement the GA.

- **P1mysenti; P2textblob; P3sentistrength; P4affin:** Weight given to sentiment analysis tool MySentiment API, Textblob, SentiStrength and Affin respectively. This weight is characterized by a variable of type integer that ranges between 1 and 4. The weight assigned to each tool means the importance that every tool has in the course of the simulation. For example, assuming randomly MySenti tool has a weight of 4, 3 for TextBlob, 2 for SentiStrength and 1 for Affin, this means that during the simulation the values of the daily scores calculated by MySentiment Api will have a greater influence than the other three and so on.

- **PosLevel:** Double variable type that defines the positive minimum level required to make a purchase order. Every day, for each company, is read a total score equivalent to the positive or negative score off the day. Randomly the value attributed to this gene is the minimum value that the total score of the day has to have to be made a purchase that day. This gene may contain values between 500 and 10000.

- **NegLevel:** Double variable type that defines the negative minimum level required to make a sales order. This gene is used to find the threshold value for an output method. One of the implemented output methods, closed position when it finds a negative peak score (output method by score). In this case, this gene is assigned the minimum limit value is required to have in the total daily score for closing position. This gene may contain values between -100 and -10000.

- **Days:** Number of days to keep the stocks bought on hold. It is an Integer variable that varies between 20 and 60. That is at least buy stocks of a company and we waited 20 days to sell or so at most buy stocks and sell past 60 days.

- **NStocks:** This gene is a variable of type Integer that ranges between 5 and 10. In other words is the maximum number of stocks that we have in my portfolio.

- **StopTwitter:** This gene is an integer variable that can have only two values, zero or one. Depending on the value attributed is an output method to the algorithm.
  StopTwitter = 0: Uses in the strategy the output methods for *time* and *price*.
  StopTwitter = 1: Uses in the strategy the output methods for *time*, *price* and *score*.
  In both cases the output method which occurs first is the method that closes the position.

- **StopGain:** This gene is responsible for giving a sales order, depending on the value attributed to the gene. This sales order is intended to allow the gain set targets are qualified. This variable is of type Double and can get values between 15 and 50. For example if you make a purchase of a stock of €100.00. Imagining randomly gene is assigned the value 25. Then is given a winning bar from the date of purchase of the stock that defines a return of 25%. That is, when the stock has the value of €125.00 we sell stock.

- **StopLoss:** This gene is responsible for giving a sales order, depending on the value attributed to the gene. This sales order aims to prevent losses. This variable is of type Double and can get values between 5 and 20. For example if you make a purchase of a stock of €100.00. Imagining randomly gene is assigned the value 10. Then is given a maximum loss from the date of purchase of stock. That is, when the action gets to €90.00 we sell stock.

➢ *Fitness Function*

In genetic algorithms the individuals (chromosomes) are evaluated according to the fitness function, which defines the problem under study, ie, the fitness function provides a measure of how individuals behave in the field of the problem. This function has a very important role in the implementation of the GA to obtain the best solution within a large number of solutions. Good fitness functions help GA to explore the search space more effectively and efficiently.

Below is presented a detailed description of the fitness function implemented in order to overcome the portfolio optimization problem.

Implementation of Fitness Function:

Initially, the program loads all asset prices and the scores calculated for each of the four sentiment analysis tools for each of the thirty companies during the training period.

Once uploaded all the files, the days of our training period begin to be analyzed. Daily we calculate the total score value through the first four genes that define the weights of the tools. If the calculated total value exceeds the value Pos Level (gene 5) and the portfolio has space (gene 8), the algorithm makes up a purchase. The exit point depends of StopTwitter (gene 9), explained in more detail below in investment rules section. If the Stop Twitter is zero, the exit point is the first to happen over time of the following genes: over time (gene 7), stop gain (gene 10) and through the stop loss (gene 11). On the other hand, if the StopTwitter is one, the purchase is subject to the three methods mentioned above and yet another method that is per score. That is, if it happens a peak of very negative score that exceeds the NegLevel (gene 6) before the days pass the position is closed. In case all output methods, the first to occur is the method which will close the position. Through this system, the stocks are bought to the portfolio until the NStocks limit is reached (gene 8). When the training period ends a new population is generated with the best chromosome and new chromosomes resulting from crossovers. The process is repeated with the new population until the number of evolutions / generations is achieved.

For the implementation of the fitness function that the chromosomes will undergo during the evolutionary process resorted to some types of output methods ie investment rules that we will explain below.

Investment Rules:

For the development of the fitness function were implemented some investment rules, depending on the order of buy or sell, that will be performing.

The algorithm developed calculates scores through the four sentiment analysis tools implemented and whenever a positive event is detected, or when a daily peak with a very positive score, a purchase order is generated.

The algorithm sales order is generated whenever one of the following combinations occurs firstly:

1. **Time:** where is defined a variable set of days to wait to close position - *Gene 7*;
2. **Score:** the position is closed if a very negative event was detected. The event is only detected if the daily score presents, at a minimum, the value attributed to *Gene 6*.

3. **Price:** a positive threshold and a negative threshold for the price are set, stop gain (value attributed to *Gene 10*) and stop loss (value attributed to *Gene 11*) respectively, which are defined by a percentage of the purchase price that exceeded when the position is closed with profit or loss.

In **Figure 9** the reader can see an example where the outputs of the first two methods are used simultaneously, the method for time and score. After the time-position opening the output method waits 50 days to close the position (orange dashed line), the method score close position whenever it finds a very negative peak that exceeds the defined negative Level (blue dashed line). In the case of **Figure 9** may be seen that the position is closed by the score output method since it had a very negative peak score before the 50 days passed. If there were no such peak, the method for closing position would be the day.

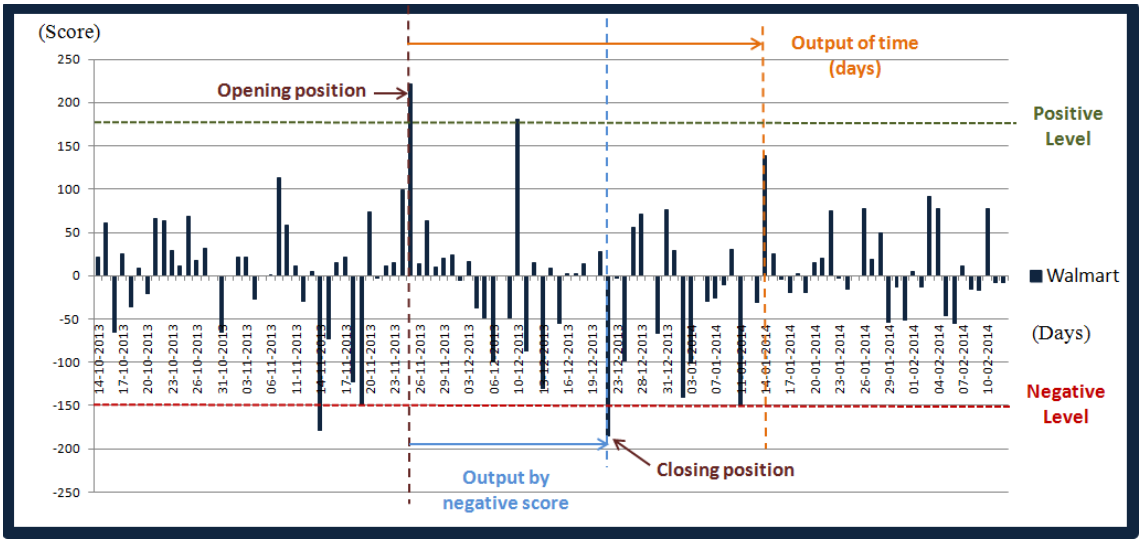

Figure 9 - Output method by time and score.

Figure 10 shows the output method price, which defines the limits of positive price (green line) and negative (red line) that if exceeded the position is closed. The position is closed with the output method that occurs first. In a real case, this may not be the best exit option as we can see in **Figure 10** where the position is closed by price and the case was closed for days (50 days) had a sale with higher gain.

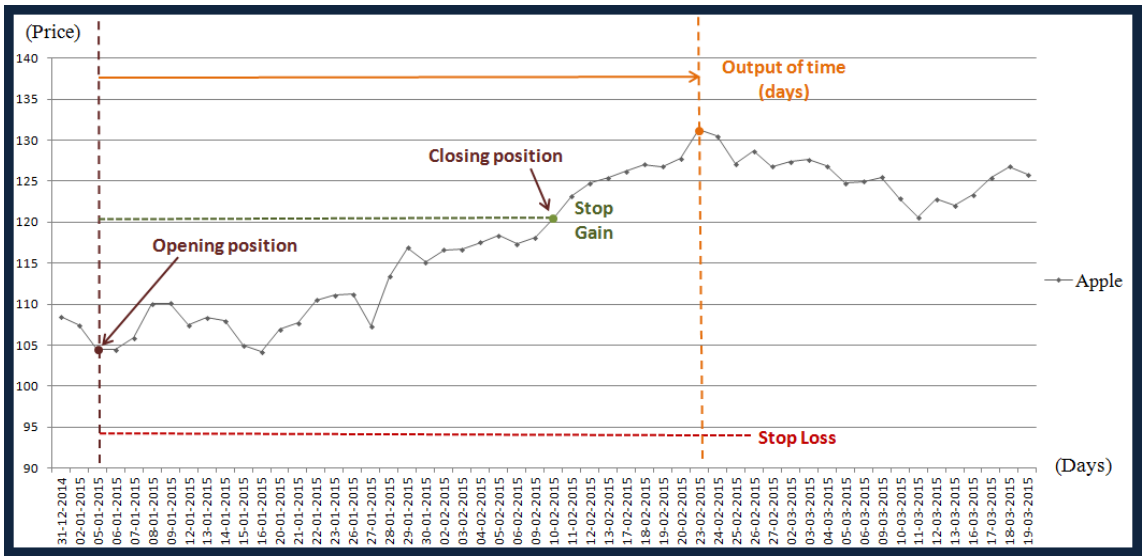

Figure 10 - Output method by price.

➢ *Selection – The New Generations*

The selection phase is an important part the implementation of the algorithm. The selection filters the best individuals in the population for the reproduction process by limiting the increasing emergence of weaker individuals. Thus, the selection should be based not only on the choice of the fittest individual, because there is a probability of less able individuals having genetic properties favorable to generation of a chromosome having the best solution to the examined problem. There are several forms of selection, the *roulette selection method*, the *selection method for the tournament* and the *selection method by "ranking"*.

The method used in this algorithm is the method of "ranking". In this method the selection is made through the sorting of individuals of the population that did better results in the fitness function, ie individuals that have obtained better returns.

➢ *Genetic Operators*

Genetic operators aim to achieve changes in the defined population, making every new generation, more capable individuals are created, thus contributing to the populations evolve with each new generation. With this, the genetic operators are classified as: crossover and mutation.

A) Crossover

This operator is considered to be the predominant genetic operator. By crossing are created new individuals mixing characteristics of two individuals "parents", **Figure 11**. This mixture is trying to imitate (at a high level of abstraction) the reproduction of genes in cells. Fragments of the characteristics of an individual are exchanged for the equivalent fragment of the others. The result of this operation is an individual who potentially combine the best features of individuals used as a base.

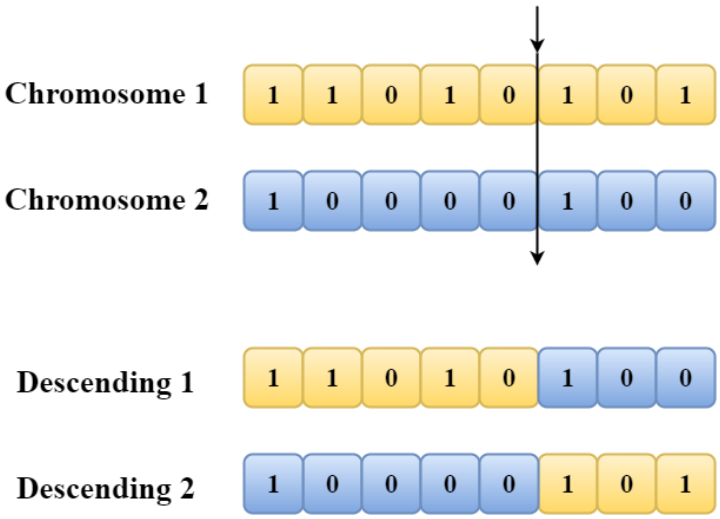

Figure 11 - Generic crossover example.

B) Mutation

This operation simply randomly modifies some characteristics of the individual on which it is applied (see **Figure 12**). This replacement is important because ultimately create new characteristic values that do not exist or appear in small quantities in the population under analysis. The mutation operator is required for the introduction and maintenance of the genetic diversity of the population. The mutation operator is applied to individuals at a generally low mutation rate.

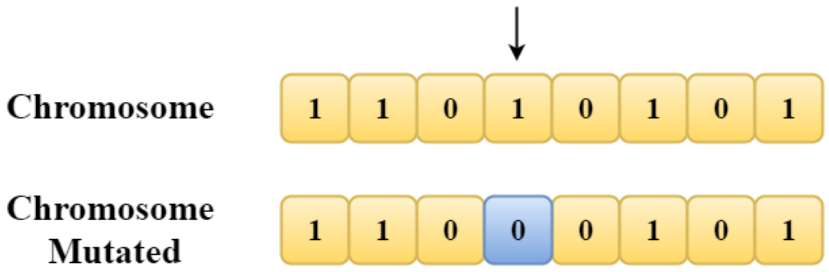

Figure 12 – Simple Mutation example.

➢ *Actualization*

At each step, a new set of individuals is generated from the previous population. At this point the individuals resulting from the crossover and mutation process are entered in the population according to the policy adopted by the genetic algorithm thus creating a new generation. In the most traditional form of the GA population maintains a fixed size and individuals are created in same number as their predecessors and replace the previous individuals altogether. It is by creating a lot of generations it is possible to obtain results of genetic algorithms.

➢ *Finalization*

The finalization operator is responsible for determining whether the implementation of Genetic Algorithm (Evolution of the population) will be completed or not. Such action is carried out from the execution of tests based on a predetermined stop condition. This stop condition may vary from the amount generations developed to the depth of a fitness value of each chromosome in a population. In this implementation the stop condition is defined by number of generations.

Chapter 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 11 user accounts set initially. The extracted data is from September 2013 to September 2015.

This chapter is divided into four main parts. Section 4.1 presents the details of the data collected. Section 4.2 presents the development and validation of the classification model of event detection. In this section we present three case studies that demonstrate the detection of events in companies. In Section 4.3 we present the case study that shows the results of the best solutions of the genetic algorithm. Finally, Section 4.4 presents the case study that analyzes the popularity of companies and presents the best strategy based on popularity for returns.

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 5** shows some relevant information in the data collection stage to the completion of the project.

Table 5 - 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 of Event Detection

This section will present three case studies that validate the event detection. Initially we present a case study about the Apple's Company, then present a case study of Microsoft and finally a case of Walmart. In these case studies we detect important events in the lives of companies and present examples of tweets that confirm the events detected.

4.2.1. Apple – Case Study I

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 **Figure 13** 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.

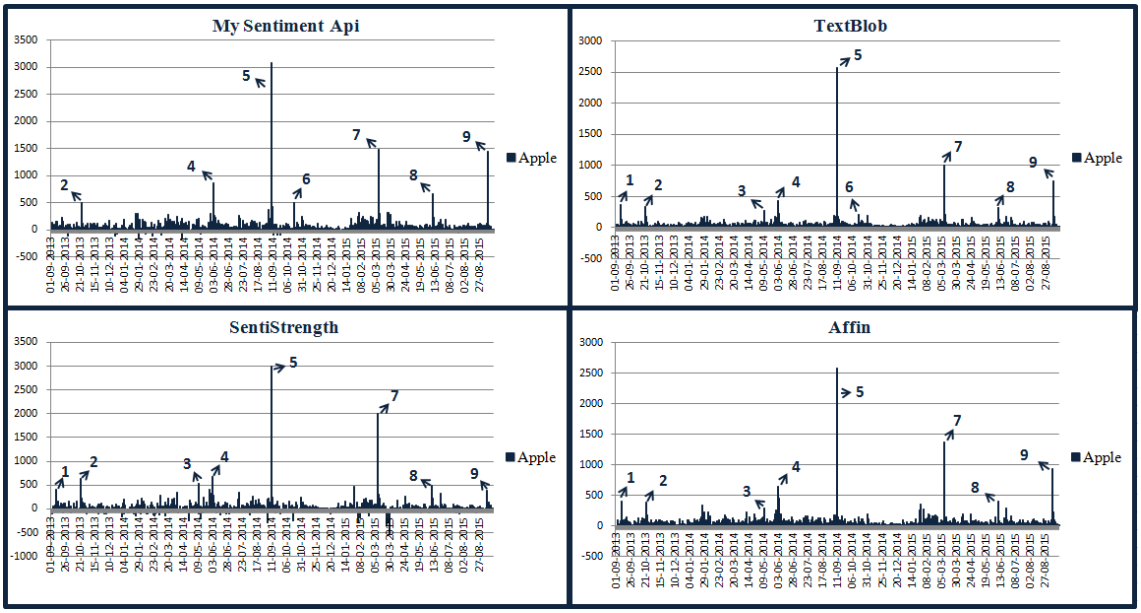

Figure 13 - Sentiment analysis over time – Apple's Company.

The special event for each number in this graph is indicated in **Table 6**. 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 6** are examples of tweets analyzed on days of events in order to confirm that users published information about the special events.

Table 6 – Examples of tweets for Apple Special 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	Apple launches the	"Happening now. Apple's #WWDC2015 Christmas in June.

(8)	big WWDC	Boom.http://t.co/Tl8p98gv5b 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 6** 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 6**, 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.

4.2.2. Microsoft – Case Study II

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 **Figure 14**, shows the graphs for each sentiment tool used.

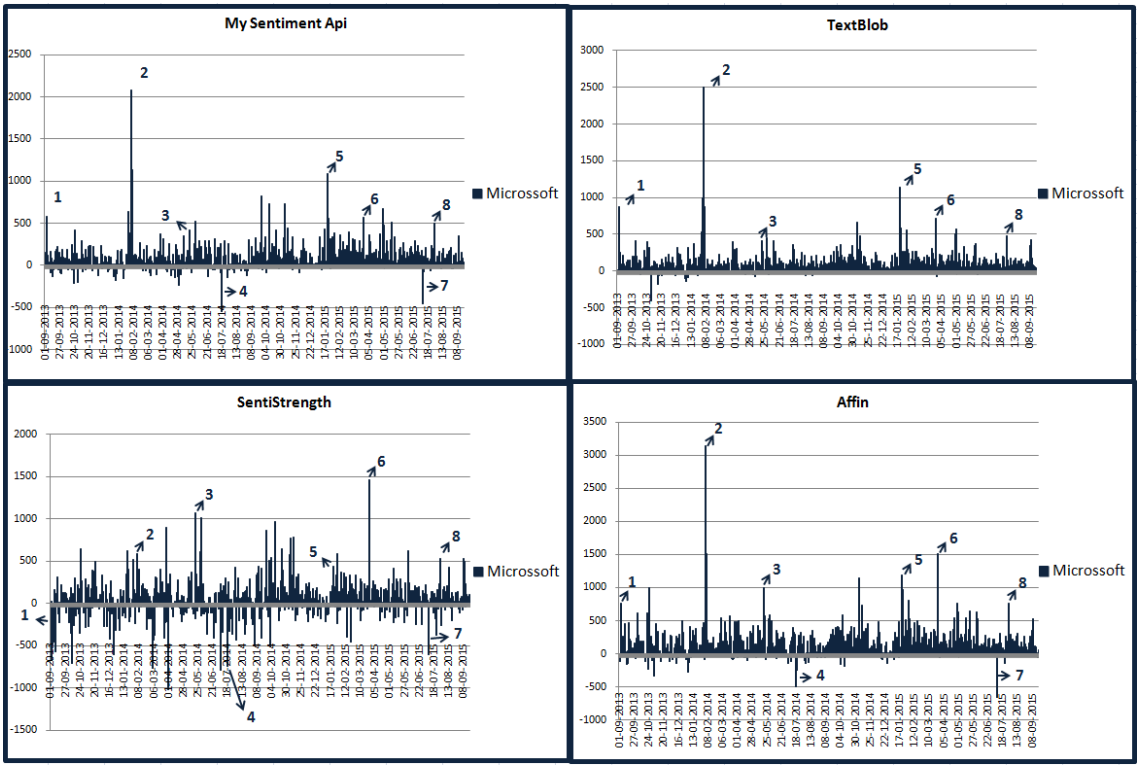

Figure 14 - Sentiment analysis over time – Microsoft's Company.

The first event highlighted in the **Figure 14**, 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 7** 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 7** 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 7 - 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	**Microsoft**	"Microsoft makes cheaper version of Surface Pro 3, with smaller

(6)	**launches Surface Pro 3**	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 7**, the existing tweets in our data collection correspond to the detected content in the events.

4.2.3. Walmart – Case Study III

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**. Figure 15** shows the graphs corresponding to the sentiment evaluation of the company for each of the four tools used in the test period.

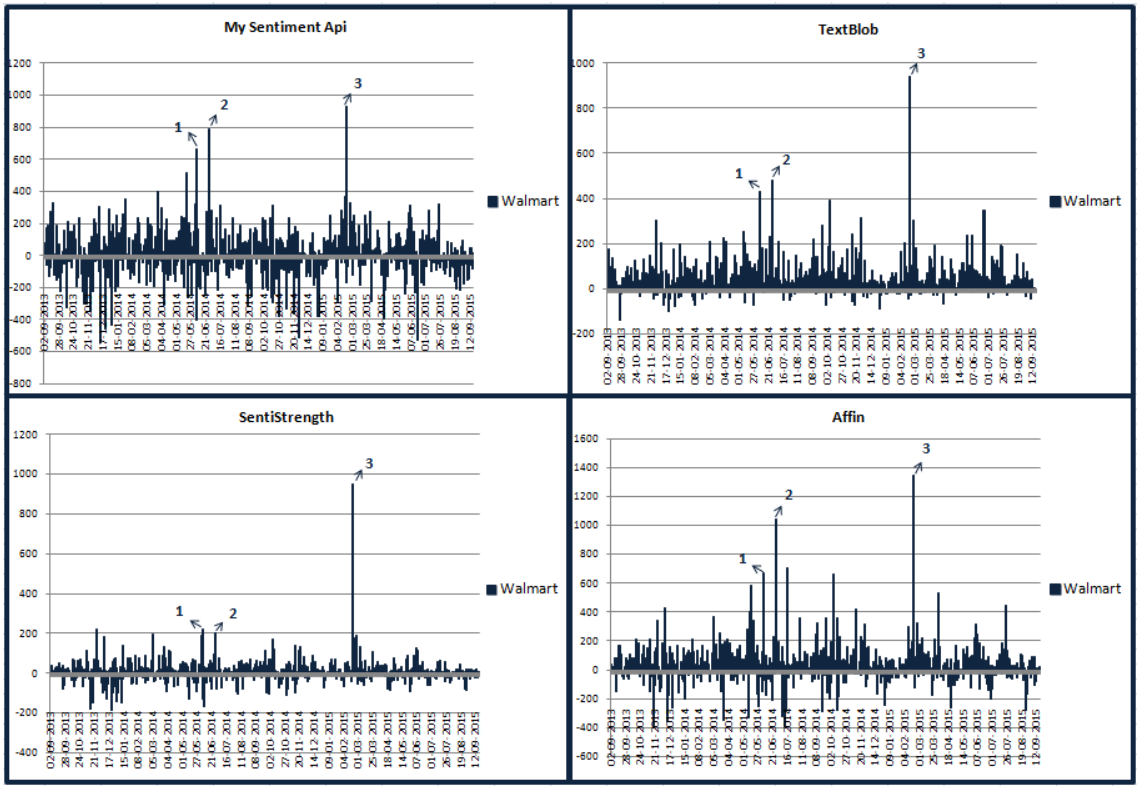

Figure 15 - Sentiment Analysis over time – Walmart's Company.

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 8**.

Another event that has not been indifferent to the financial community proposed here was the first event detected in **Figure 15**. The Annual Shareholders' Meeting is an annual event held on the day 6 June 2014. In **Table 8** 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 8** 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 8 - 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.

4.3 Development and validation of classification model of Genetic Algorithm

After identifying the optimal strategy for the training period September 2013 to September 2014 on the Dow Jones index the algorithm was applied to a real scenario for the period September 2014 to September 2015. That is, the training of the solutions is done for one year and the next year is carried out the test with the best solutions of the previous year, thus ensuring that the best solutions found do not know the market behavior in the following year.

The initial parameters of the genetic algorithm passed through an initial population of 128 individuals and 50 generations as stop condition. Training tests were repeated 5 times and the results in **Table 9** show the average of five best results obtained in the test period. As already mentioned, **Table 9** shows the results of the best solutions obtained by the algorithm implemented, but also presents the results of B&H strategy to compare. B&H is a passive investment strategy in which an investor buys stocks and holds them for a long period of time, regardless of fluctuations in the market. An investor who employs a buy-and-hold strategy actively selects stocks, but once in a position, is not concerned with short-term price movements and technical indicators. This strategy, B&H, will serve as a point of comparison with our strategy implemented over this case study.

Table 9 - Results of the average investment strategies in test year.

Year	Nº operations	Success rate	Average days in the market	GA Return			B&H Return
				Worst	Average	Best	
Sep. 2014 To Sep. 2015	45	47,13%	47	-3.73%	-1.95 %	0.25%	-4.2%

The table above shows the results in the real test period between September 2014 and September 2015. The number of operations is the average value of transactions performed in the five best chromosomes. The success rate was 47.13%. This number represents the average number of times that a trade obtained a profit in the total number of operations. The average days in the market were 47. That is, after opening a position we closed it 47 days later. For the five best solutions, the worst case showed a return of -3.73%, the best case presented 0.25% and the average return of the five solutions was -1.95%. These results are better than the B&H strategy that returned -4.2%.These results were compared with the Buy & Hold strategy in terms of return for the same period as can be seen in the **Figure 16**.

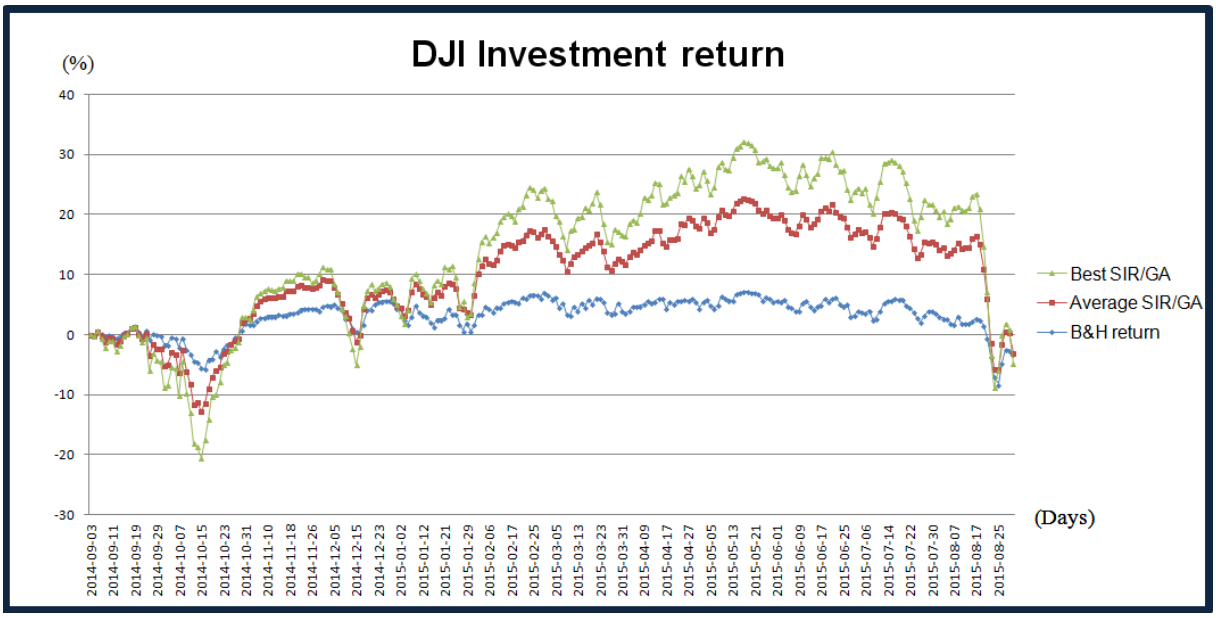

Figure 16 - DJI return of different strategies compared with B&H.

The figure above shows good results in the context that the line of our best solution (green line) and the average line of the five best solutions (red line) are above the line that is the strategy B&H (blue line). This means that over the year, by the implemented algorithm we have better results than the return of DJIA. In the analysis of **Figure 16** there are two critical moments that our strategy was below expectations showing great volatility. These volatility moments happen in most cases influenced by external factors that may be news about the market in which the companies operate or to political or economic incidents.

For example, the event that occurred in mid-August 2015 and oscillated the market with a big drop in the stock market was difficult to predict. These days were characterized by the sharp devaluation of the Yuan which has raised fears of a slowdown in the Chinese economy. This devaluation had a strong impact on the market and created a storm that devastated many emerging markets. From Asia to Latin America, many emerging countries have registered strong devaluations in the currency. A trend that was not new but in August 2015 China was even stronger.

The developed algorithm focuses on popular companies in the Dow Jones, or in companies that were detected special events published on Twitter by the financial community. The decline in the week after August 22, 2015 was characterized by a big decrease on the best performing stocks in the previous year. Nevertheless we obtained results on average higher than the index return and if we had stayed out of the market in these two events is expected that the results would increase greatly.

➢ *Real Test – Best Solution*

Our test year was not a good 12 month period in terms of returns since the performance of the Dow Jones index was -4.2%. In this case the prices followed the trend that can be seen in **Figure 17**.

Figure 17 - DJI index performance from Sep 2014 to Sep 2015.

The investment strategy that had the best result this year had a return 0.25%. The values assigned to more capable chromosome are defined in **Figure 18**. It can be seen in the figure that the waiting period to close position by the method by time got a standby time of 44 days. The gene StopTwitter has a value of zero in this strategy meaning that the output method through score was not used. In this strategy the algorithm performed 48 trades, 14 operations closed by price (stop loss) and the rest operations closed by time, which was a period of 44 days. This investment strategy had 21 successful operations, which resulted in positive profit in 48 possible operations. It can be seen that a positive result has been achieved even if the number of successful cases has been smaller than the case of failure. In **Table 10** is presented as an example the trades for the period between September 3, 2014 and December 26, 2015. It is important to note that the exercise started with an initial investment of EUR 10 000 and the amount invested whenever we open a position is divided equally by the number of stocks (Gene 8). That is, as the number of stocks is 5 (see **Figure 18**) and the initial budget is EUR 10 000, will initially invest EUR 2000 in each stock.

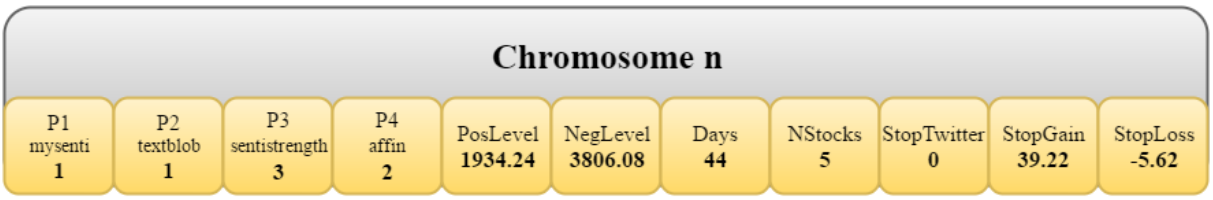

Figure 18 - Values assigned to chromosome that is the best solution.

Table 10 - Example of transactions on the best strategy.

Company	Buy	Number of Stocks	Investment	Sell (Time)	Sell (Stop Loss)	Profit
Chevron	03-09-2014	16	1939.03	17-10-2014	29-09-2014	-110.85
Exxon Mobil	04-09-2014	21	1982.74	20-10-2014	09-10-2014	-131.83
Merck	08-09-2014	34	1996.29	22-10-2014	14-10-2014	-140.90
3M	08-09-2014	14	1965.88	22-10-2014	10-10-2014	-150.06
Apple	09-09-2014	20	1918.65	23-10-2014	-	133.93
General Electric	29-09-2014	79	1937.36	12-11-2014	-	83.84
Johnson & Johnson	09-10-2014	19	1871.50	24-11-2014	-	100.76
DuPont	14-10-2014	30	1817.29	28-11-2014	-	165.43
3M	15-10-2014	14	1821.10	28-11-2014	-	362.66
Exxon Mobil	23-10-2014	21	1897.06	08-12-2014	-	-35.24
United Health	12-11-2014	21	1978.09	26-12-2014	-	150.02
Nike	24-11-2014	21	2034.39	07-01-2015	-	-58.30
Johnson & Johnson	01-12-2014	19	1993.49	14-01-2015	-	-74.36
Caterpillar	03-12-2014	21	2027.92	16-01-2015	10-12-2014	-144.94
Merck	08-12-2014	34	2038.19	21-01-2015	12-12-2014	-122.65
Pfizer	15-12-2014	66	1970.95	28-01-2015	-	69.61
Merck	16-12-2014	35	1940.45	29-01-2015	-	180.72
Disney	26-12-2014	21	1984.09	09-02-2015	-	139.88

4.4 Development and validation of classification model of Popularity

Based on the last four case studies, we decided to make one last case study to try to understand if the popularity of each company is a strong indicator to predict the returns of DJIA. This case study analyzes the popularity of each company between September 2013 to September 2014 and with the popularity results we will decide the companies to invest in the next year. The popularity is detected based on four sentiment analysis tools separately. That is, in this case study we will only take into account the volume of positive tweets detected by sentiment analysis tools.

Initially with the 30 companies in the Portfolio, was performed a count of the number of positive tweets for each company. In Table 11 we present the 15 companies that have obtained the highest number of positive tweets during the training period with the respective volume for each tool.

Table 11 - Fifteen more popular companies calculated by the four sentiment analysis tools.

MY SENTIMENT API		TEXTBLOB		SENTISTRENGTH		AFFIN	
COMPANY	VOLUME OF TWEETS	COMPANY	VOLUME OF TWEETS	COMPANY	VOLUME OF TWEETS	COMPANY	VOLUME OF TWEETS
Apple	8082	Apple	5592	Apple	4918	Apple	4957
Intel	3872	Intel	2641	Intel	2584	Intel	2274
Microsoft	2063	Microsoft	1294	Microsoft	1145	Microsoft	1155
Cisco	1867	Cisco	1208	Cisco	1024	Cisco	1098
IBM	1605	IBM	1009	IBM	907	IBM	1010
Disney	1259	Disney	815	Disney	714	Disney	827
Visa	1054	Walmart	712	Walmart	707	Visa	654
Walmart	993	Nike	587	Nike	698	Walmart	608
Nike	961	Visa	563	Visa	554	Nike	559
Coca-Cola	735	Coca-Cola	422	Verizon	376	Coca-Cola	412
Boeing	384	Verizon	286	Coca-Cola	276	Verizon	270
Verizon	380	Boeing	277	Boeing	268	DuPont	262
DuPont	366	DuPont	247	DuPont	243	Boeing	261
McDonalds	305	McDonalds	220	McDonalds	206	Goldman Sachs	175

Goldman Sachs	284	Goldman Sachs	169	Goldman Sachs	157	McDonalds	157

That said, we calculated the annual return if we invest in the fifteen most popular companies. Our strategy is to invest in the most popular companies in the previous year.

The objective of this study was to try to understand the correlation between the popularity of companies and the return obtained. For this we invested between September 2014 and September 2015 in the most popular companies between September 2013 and September 2014.

Through **Table 11** is perceptible realize that the fifteen most popular companies are common to the four sentiment analysis tools. The volume of positive tweets calculated for each tool varies from tool to tool but always keeps in the same range of values. Even with the volume of tweets showing some differences and the order of fifteen companies varies it is clear that the fifteen favorite companies are the same managed to prove the consistency of the four tools implemented. Then in **Table 12** shows the return obtained by the four sentiment analysis tools. Since the fifteen companies will invest throughout the year are the same then the four tools will get the same return values when we invest in the same period.

Table 12 - Returns performed by fifteen most popular companies.

SENTIMENT ANALYSIS TOOLS	RETURN (%)
My Sentiment API	0.88
TextBlob	0.88
SentiStrength	0.88
Affin	0.88

As the four sentiment analysis tools have high consistency in relation to evaluate the volume of positive tweets, then was performed another case study. In this case study instead of investing in the fifteen most popular companies was decided to minimize the number of companies to invest during the year.

Table 13 presents the six most popular companies throughout the year and we will invest in this new case study.

Table 13 - Six most popular companies in the four sentiment analysis tools.

My Sentiment Api		TextBlob		SentiStrength		Affin	
Company	Volume of Tweets	Company	Volume of Tweets	Company	Volume of Tweets	Company	Volume of Tweets
Apple	8082	Apple	5592	Apple	4918	Apple	4957
Intel	3872	Intel	2641	Intel	2584	Intel	2274
Microsoft	2063	Microsoft	1294	Microsoft	1145	Microsoft	1155
Cisco	1867	Cisco	1208	Cisco	1024	Cisco	1098
IBM	1605	IBM	1009	IBM	907	IBM	1010
Disney	1259	Disney	815	Disney	714	Disney	827

As we previously analyze, six companies have always been in the fifteen most popular for the four tools implemented. By investing only in these six companies, we get a return of 8.13% (see **Table 14**). We conclude with this case study that the popularity of each company by volume of tweets over the year gave positive returns for the following year.

Table 14 - Return obtained by six most popular companies.

Return obtained by the Popularity Analysis (%)
8.13

Chapter 5 Conclusions and Future Work

5.1 Conclusion

In this work we present a significant evidence of dependence between stock price returns and sentiment in tweets posted about the companies. There is a signal that worth investigating which connects social networks and market behavior. But this dependence between the tweets and market behavior is only credible when tweets are properly selected.

For this reason, this thesis focuses on the development of a system capable of predicting stock returns of the Dow Jones index based on tweets posted by a financial community. In a first stage of this thesis it was implemented a software in order to detect special events in the life of the companies. Based on the detected events it was implemented a genetic algorithm to predict the evolution of the Dow Jones index. Finally, it was analyzed the possibility of the popularity of each company be a good indicator to predict the market and get positive returns.

This thesis presents a new sentiment analysis tool and uses three known sentiment analysis tools to evaluate the extracted tweets between September 2013 and September 2015. This proposed system had to deal with large amounts of data with a high level of noise typical of social networks.

With the results obtained, it can be concluded that the model detected the special events of companies' life efficiently. The system reached consistency between the four sentiment analysis tools receiving in most cases, the same events. The results demonstrated the good performance of the model indicating that the financial community is influential with respect to the publication of important tweets in life of companies.

These detected events were used to implement a genetic algorithm that had a return of 0.25% for the real test between September 2014 and September 2015. This return obtained showed an above average value when compared to the application of the Buy&Hold strategy to Dow Jones that had a return of –4.2% for the same period.

The last study in this thesis was to analyze the possibility of popularity of each company on Twitter be a good indicator to predict the market return. To analyze the popularity of each company the algorithm must consider the volume of positive tweets throughout the test period for each company. In an initial test, we tested the four sentiment analysis tools separately. For each tool calculated the total volume of positive tweets for each company and decided to invest in the fifteen most popular companies. The results showed that when applying the MySentiment Api and TextBlob tool gave the same values as when applying SentiStrength and Affin tool, 0.88% and 5.75% respectively.

The last test aimed to choose only those companies that are in the top 15 of the four sentiment analysis tools. The results showed a return of 8.13%, which allows us to conclude that the popularity of each company can be a strong indicator when the objective is to predict the market return.

5.2 Future Work

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

- ✓ The definition of a financial community is a very important step in this work to get the tweets with relevant content on the financial market. Today there are various applications and techniques that allow any user to increase the number of followers on social networks without their influence is great. This can reduce the quality of users chosen by our financial community, as the implemented method was based on the number of followers that our eleven main users had. Hence, one of the recommended improvements is try to identify the features that a user has to be on twitter to be considered influential in the stock market.
- ✓ 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. It would be interesting to take greater importance to the evaluation of retweets than just the tweets.
- ✓ 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.
- ✓ 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.

Chapter 6 References

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

Bloomberg, M. (1, October 1981). *Bloomberg*. Retrieved January 2016, from Trending on Twitter: Social Sentiment Analytics: http://www.bloomberg.com/company/announcements/trending-on-twitter-social-sentiment-analytics/

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

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.

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

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life.* London: John Murray.

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.

Freitas, A. A. (2003). *A survey of evolutionary algorithms for data mining and knowledge discovery.* Springer.

Geva, T., & Zahavi, J. (2013). Empirical evaluation of an automated intraday stock recommendation. *Decision Support Systems*.

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

Goldberg, D. E. (1989). *Genetic algorithms in search, optimization and machine learning.* Alabama: Addison Wesley.

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.

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.

Hill, M., & Liu, B. (2004). Mining and Summarizing Customer Reviews. *ACM SIGKDD International Conference on Knowledge Discovery.* Seattle, Washington.

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

Holland, J. (1975). *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence.* The University of Michigan Press.

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

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

Java Genetic Algorithms Package. (n.d.). Retrieved January 2016, from http://jgap.sourceforge.net/

Kaplan, A., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business horizons, 53(1)*, 59-68.

Kim, K.-j., & Han, I. (2000). Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index. *Expert systems with Applications, 19(2)*, 125-132.

Kisling, W., Lam, E., & Mehta, N. (2013). Human beats machine this time as fake report roils stocks. *Bloomberg News*.

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

Liu, B. (2012). *Sentiment analysis and opinion mining.* Morgan & Claypool Publishers.

Loria, S. (n.d.). *TextBlob: Simplified Text Processing*. Retrieved September 2015, from https://textblob.readthedocs.org/en/dev/

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

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?*

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

Mitchell, M. (1997). An introduction to genetic algorithms. Cambridge: Mit Press.

Mittal, Anshul; Goel, Arpit. (2012). Stock Prediction Using Twitter Sentiment Analysis.

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 September 2015, from https://finnaarupnielsen.wordpress.com/2011/03/16/afinn-a-new-word-list-for-sentiment-analysis/

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval, 2*, 1-135.

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 January 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.*

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*.

Search API. (n.d.). Retrieved January 2015, from https://dev.twitter.com/rest/public/search

SentiStrength. (n.d.). Retrieved September 2015, 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.

Simões, A., Neves, R., & Horta, N. (2010). An Innovative GA Optimized Investment Strategy based on a New Technical Indicator using Multiple MAs.

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. *Information Systems for Crisis Response and Management Conference.* Seattle, WA, USA.

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

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.

Thakkar, R. (2007). Data mining in Stock Market. *Asia Pacific Journal of Research, I Issue IX*.

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.

Tsytsarau, M., & Palpanas, T. (2012). *Survey on mining subjective data on the web.*

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*, 376-378.

Yamamoto, Y. (n.d.). *Twitter4J*. Retrieved 2015, from http://twitter4j.org/en/index.html

Yang, Steve Y.; Mo, Sheung Yin Kevin; Zhu, Xiaodi. (2013). An Empirical Study of the Financial Community Network on Twitter.