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 events of real life. 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 (DJIA), Twitter, Sentiment Analysis, Event Detection, Genetic Algorithm

I. INTRODUCTION

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

Several techniques are known and applied by investors in order to increase profit and minimize risk. One of the most used is Technical Analysis that is based on the hypothesis that past patterns that relate to behavior of prices in each asset tends to repeat itself in the future [1]. 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 [2]. It is also through the results of the Fundamental Analysis that can be determined whether an asset is overvalued or undervalued.

The greater or lesser supply/demand of a certain stock can be influenced by the historical behavior of prices, future prospects related to the performance of the company issuing of stock or even by news published in blogs, social networks. Studies by [3], [4], [5] and [6] 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 is positive or negative and can in some way influence financial markets.

Today, the Internet is used to find information but also to share information, share knowledge and also serve as a channel for business. Online social networks are part of millions of people worldwide every day, becoming over the years an important communication platform that brings together various information, including opinions and feelings expressed by users in sharing content in news messages published. Social networks have attracted the attention of many researchers who aim to correlate the content of the various publications with the events of real life [7]. The large interest happens because many events are published by the traditional media with a delay, while on social networks the event is published almost as instantly. Thus, the big question is to what extent the information provided on social networks truthfully reflects the actual events and is it possible to use this information to detect events.

The objective of this work is to develop a system 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, extraction of tweets published by the financial community, the analysis of the content of tweets, the detection of special events and the implementation of a genetic algorithm in order to present optimal solutions to the problem of predicting stock returns. In this methodology, the actual events we are looking for are events that somehow influenced positively or negatively the financial community, i.e events that are retweeted in the tweets published with implicit feelings about companies. Examples of such events: product launches, special events in the life of companies, events, campaigns, among others.

This project presents an original work with respect to event detection, implementation of an evolutionary algorithm and study of the popularity of companies. There are seven main contributions. First the algorithm that defines a financial community inside Twitter and extracts all tweets for each user. Then the tweets filtering algorithms that allow us to get only the tweets related to the financial market more properly the thirty companies that compose the Dow Jones index. The
third contribution of this paper is the creation of a simple sentiment analysis tools and the conjunction with other three
known sentiment analysis algorithms. Through the four algorithms we evaluate the tweets individually in positive,
negative and neutral scores. The fifth contribution of this
paper is the manual detection of events that have been
characterized by large positive peaks and large negative
peaks and are directly related to special events that positively
and negatively affected the financial community. The last
contribution is the creation of investment strategies to predict
the trend of the Dow Jones index through implicit sentiment
in tweets with implementing a genetic algorithm and
analyzing the popularity of each company.

This paper is organized as follows: Section 2 presents the
state-of-the-art for some concepts relating the use of genetic
algorithms in order to predict the market, some tools of
sentiment analysis and finally presents some literature
concerning the social network Twitter, their influence on the
market and still some works produced in the area of event
detection on Twitter; Section 3 presenting the system
architecture and the detailed explanation of each module;
Section 4 discusses the experiments and results. Finally, the
conclusions are drawn in Section 5.

II. RELATED WORK

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 [8]. 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 in 1960 by John Holland [9], with the
initial objective to study phenomena related to the adaptation
of species and natural selection that occurs in nature [10], as
well as develop a way to incorporate these concepts to the
computers [11]. Many studies have been done in the area of
GAs in order to implement new investment strategies for the
stock market. A recent article [12] proposed a new
investment strategy focusing on the S&P500, FTSE100,
DAX30 and the Nikkei225 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 crossover (MACs) and
derivatives Moving Average (MAD)).

The use of genetic algorithms, neural networks and genetic
programming in an attempt to find a low cost solution is very
common. But in order 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
relevance to an increasingly exploited technique, Sentiment
Analysis. 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. [13] released
ANEW, a lexicon with affective norms for English words.
The application of ANEW to Twitter was explored by [14],
leveraging the AFINN lexicon. [15] 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 [16],
leveraging the SentiStrength. [17] and later [18] extended the
well known Wordnet lexical database [19] 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.

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, [20] 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, [21] 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 to spread
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 Associated 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 contrarily to humans, who would have realized that the information was false before trading on it, [22]. 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.

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, [23] published a study that consists of forming a financial community on Twitter where users of this community share interests in the financial market. The results illustrate that a strong interdependence between the social climate and the movement of stock prices by creating the financial community can be created. This study could be concluded that the sentiment generated by each node of the financial community has predictive power in consistent market returns and own market volatility. 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 as 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. The detection of natural disasters and social events using the social network Twitter has been widely analyzed and discussed by researchers. [24] and [25], 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 [26] and [27].

With 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.

III. METHODOLOGY

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 following.

3.1. TWITTER

A) Definition of a Financial Community

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

B) 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 in eleven persons that are very known in the financial community. But the final objective is to find the persons that follow them. The second stage of selection of user accounts was based on the followers of this community of eleven.
users. There is a strong likelihood that these followers belong to the same financial investment community thus sharing similar interests. Yet there was a concern about the selected followers. Only the second stage users who have more followers than users that they follow were chosen. In Figure 2 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.

3.2. TWEETS FILTER
With the financial community defined and all tweets for each id extracted, we implemented software to filter tweets to eliminate useless content for our project. The filtering is illustrated in Figure 3. 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. 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 organizes tweets in thirty files where each file corresponds to one of thirty companies in the Dow Jones Average. For a tweet to match one of the companies it just takes one of the keywords listed in file. 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.

3.3. DEFINING SENTIMENT
It is often unclear whether a tweet contains some feeling. In this study, we used four text analysis tools to assess this feeling often difficult to find. One of the tools used was developed by us and the rest were developed by different authors. Next are described briefly and in Table 1 is an example of the evaluation of two tweets to four different tools.

My Sentiment 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 [28]. 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. [29], 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 [30] 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).

Affen: Affen 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 [31] 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 1 an example of six tweets is presented with the appropriate calculation of scores for the four sentiment applications mentioned above.

<table>
<thead>
<tr>
<th>THEORETICAL ANALYSIS</th>
<th>TWEETS</th>
<th>SCORE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>MENIMENT</strong> : <strong>ART</strong></td>
<td><strong>TEXT</strong> : <strong>BIE</strong></td>
</tr>
<tr>
<td>Positive</td>
<td>“This bigger people thing by Apple makes H2O look good. Very good.”</td>
<td>2,44</td>
</tr>
<tr>
<td>Positive</td>
<td>“We live in an exciting period of transformation in Psychology. #Mindset”</td>
<td>3,32</td>
</tr>
<tr>
<td>Negative</td>
<td>“Apple stock is falling since today’s announcements. Has surprised”</td>
<td>-1</td>
</tr>
<tr>
<td>Negative</td>
<td>“Cisco is down 12% after a disappointing earnings release. #Cisco #CSCO why he thinks it’s still a buy. <a href="http://t.co/%ISEYo/CBC%E2%80%9D">http://t.co/%ISEYo/CBC”</a></td>
<td>-12</td>
</tr>
<tr>
<td>Neutral</td>
<td>“Walmart informational meeting on GB over plans for donation. @CityofGreenBay @WFRVNews #walmart <a href="http://t.co/5ywrgJ2R72%E2%80%9D">http://t.co/5ywrgJ2R72”</a></td>
<td>0,078</td>
</tr>
<tr>
<td>Neutral</td>
<td>“Catch me on CNN tomorrow at 6:40AM talking Apple’s announcements. Tune in!”</td>
<td>0,749</td>
</tr>
</tbody>
</table>

### 3.4. NORMALIZATION
Normalization of data is an important step in the system architecture. Since the use of the Twitter social network has been increasing over time it is important 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 the number of tweets increases exponentially between 2013 and 2015 is necessary to make normalization by volume of tweets, so as to achieve conformity in very positive and negative peaks. 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.

\[
\text{Score}_{\text{Normalized}} = \frac{\text{Score}_{\text{Daily}}}{\text{Total Tweets last 3 months}}
\]

To calculate the normalization for each day is necessary 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.

### 3.5. DETECT EVENTS
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.

### 3.6. GENETIC ALGORITHM
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. Figure 4 shows the structure of operation of a genetic algorithm. 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.

![Figure 4 - Structure of operation of a traditional GA.](image)

**A. 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. Figure 5 shows the chromosome model defined in this methodology. Our model chromosome consists of 11 genes, i.e 11 different variables which we will explain the function of each of them is followed.

![Figure 5 - Definition of the chromosome model used to implement the GA](image)

Basically the first four genes are responsible for assigning weights to the importance of sentiment tools. The daily scores for each tool values are multiplied by the weights and then the scores of the four tools are added in order to stay with a total score daily. The fifth gene (PosLevel) is a
random value that corresponds to the positive minimum level that the total daily score has to have to make a purchase. Thus, the sixth gene (NegLevel) also is a random value that corresponds to the negative minimum level the score total daily has to have to make a sale. The seventh gene (Days) is the gene responsible for the output method through the days. Depending on the value assigned to the gene is the number of days that is expected to close position. The eighth gene (Stocks) is the gene which defines the portfolio size, or the maximum number of open positions. The number nine gene can be zero or one and defines the combination of output methods. Finally, the last two genes are two output methods that are characterized by the output through price. The gene number 10 and 11 are responsible for giving a sales order, depending on the value attributed to the gene. In case of gene 10, this sales order is intended to allow the gain set targets are qualified and in the case of gene 11 this sales order aims to prevent losses.

B. 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). 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.

IV. 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.

4.1. Development and validation of classification model of Event Detection (Apple Case Study)

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

![Figure 6 - Sentiment Analysis over time – Apple’s Company.](image)

The special event for each number in this graph is indicated in Table 2. 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.

In Table 2 are examples of tweets analyzed on days of events in order to confirm that users published information about the special events. It is possible to observe 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 2, 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.
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 7.

4.2. 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 3 show the average of five best results obtained in the test period. As already mentioned, Table 3 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 2 - Examples of tweets for Apple Special Event.

<table>
<thead>
<tr>
<th>Date</th>
<th>Apple Special Event</th>
<th>Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>10-09-2013</td>
<td>Apple announced the iPhone 5C and</td>
<td>&quot;Apple unveiled its new iPhone 5C and iPhone 5S, multi-colored iPhone 5C <a href="http://cbsi.co/1I7uMGa">http://cbsi.co/1I7uMGa</a> via @TheTel&quot;</td>
</tr>
<tr>
<td></td>
<td>iPhone 5S</td>
<td></td>
</tr>
<tr>
<td>22-10-2013</td>
<td>Apple announced iPad Air and iPad</td>
<td>&quot;RT @Forbes: The new iPad Air and iPad Mini will maintain Apple's premium positioning in the tablet market <a href="http://cbsi.co/1I7wJMc">http://cbsi.co/1I7wJMc</a> via @TimeTech&quot;</td>
</tr>
<tr>
<td>09-07-2013</td>
<td>iPhone 6</td>
<td>&quot;RT @ReelReports24: Apple iPhone 6 got a launch date finally. Launching in August, 2014! #MobileNews #Technology #Apple&quot;</td>
</tr>
<tr>
<td>02-06-2014</td>
<td>Apple presented the new version of</td>
<td>&quot;Apple presents the iPhone 6 in August, the newspaper says&quot;</td>
</tr>
<tr>
<td></td>
<td>iOS and iPad Air</td>
<td></td>
</tr>
<tr>
<td>09-09-2014</td>
<td>iPhone 6 and iPad 6 Plus</td>
<td>&quot;Apple live blogging and live streaming Thursday event in 3 hours <a href="http://cbsi.co/1I7wJMc">http://cbsi.co/1I7wJMc</a> via @TheTel&quot;</td>
</tr>
<tr>
<td>09-05-2015</td>
<td>Apple Watch</td>
<td>&quot;Apple's new iWatch is to be unveiled at the iPhone 6 Plus. #AppleEvent&quot;</td>
</tr>
<tr>
<td>09-05-2015</td>
<td>iPhone 6</td>
<td>&quot;Apple's new iWatch is to be unveiled at the iPhone 6 Plus. #AppleEvent&quot;</td>
</tr>
<tr>
<td>08-06-2015</td>
<td>Apple Watch</td>
<td>&quot;Apple launched the Apple Watch at WWDC&quot;</td>
</tr>
</tbody>
</table>
| 09-09-2015 | iPhone 6 and iPad 6 Plus           | "Apple announces iOS 9. Tim Cook calls it a giant release."

### Table 3 - Results of the average investment strategies in test year.

<table>
<thead>
<tr>
<th>Year</th>
<th>No. operations</th>
<th>Success rate (%)</th>
<th>Average days in the market</th>
<th>SHIR/GA Return (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014 to</td>
<td>45</td>
<td>47.13</td>
<td>47</td>
<td>-3.73</td>
</tr>
<tr>
<td>2015</td>
<td></td>
<td></td>
<td></td>
<td>-1.95</td>
</tr>
</tbody>
</table>

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 DIJA. In the analysis of Figure 7 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 8.

![Figure 8 - DJI index performance from Sep 2014 to Sep 2015.](image)

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 9. 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.

![Figure 9 - Values assigned to chromosome that is the best solution.](image)

4.3. 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 4 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>
<thead>
<tr>
<th>COMPANY</th>
<th>VOLUME OF TWEETS</th>
<th>COMPANY</th>
<th>VOLUME OF TWEETS</th>
<th>COMPANY</th>
<th>VOLUME OF TWEETS</th>
<th>COMPANY</th>
<th>VOLUME OF TWEETS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apple</td>
<td>8082</td>
<td>Apple</td>
<td>5592</td>
<td>Microsoft</td>
<td>4918</td>
<td>Microsoft</td>
<td>4917</td>
</tr>
<tr>
<td>Intel</td>
<td>3872</td>
<td>Intel</td>
<td>2641</td>
<td>Pfizer</td>
<td>2584</td>
<td>Pfizer</td>
<td>2234</td>
</tr>
<tr>
<td>Microsoft</td>
<td>2063</td>
<td>Microsoft</td>
<td>1294</td>
<td>Home Depot</td>
<td>1145</td>
<td>Home Depot</td>
<td>1155</td>
</tr>
<tr>
<td>Cisco</td>
<td>1907</td>
<td>Cisco</td>
<td>1208</td>
<td>Walmart</td>
<td>1024</td>
<td>Walmart</td>
<td>1098</td>
</tr>
<tr>
<td>IBM</td>
<td>1605</td>
<td>IBM</td>
<td>1009</td>
<td>Goldman Sachs</td>
<td>907</td>
<td>Goldman Sachs</td>
<td>1010</td>
</tr>
<tr>
<td>Disney</td>
<td>1259</td>
<td>Disney</td>
<td>815</td>
<td>Travelers Companies</td>
<td>714</td>
<td>Visa</td>
<td>827</td>
</tr>
<tr>
<td>Visa</td>
<td>1054</td>
<td>Walmart</td>
<td>712</td>
<td>Merck</td>
<td>707</td>
<td>Traders Companies</td>
<td>654</td>
</tr>
<tr>
<td>Walmart</td>
<td>903</td>
<td>Nike</td>
<td>587</td>
<td>Visa</td>
<td>698</td>
<td>Merck</td>
<td>688</td>
</tr>
<tr>
<td>Nike</td>
<td>961</td>
<td>Visa</td>
<td>563</td>
<td>United Technologies</td>
<td>554</td>
<td>United Technologies</td>
<td>559</td>
</tr>
<tr>
<td>Coca-Cola</td>
<td>735</td>
<td>Coca-Cola</td>
<td>422</td>
<td>Nike</td>
<td>376</td>
<td>Nike</td>
<td>412</td>
</tr>
<tr>
<td>Boeing</td>
<td>384</td>
<td>Verizon</td>
<td>286</td>
<td>United Health</td>
<td>278</td>
<td>Cisco</td>
<td>270</td>
</tr>
<tr>
<td>Verizon</td>
<td>380</td>
<td>Boeing</td>
<td>277</td>
<td>Cisco</td>
<td>268</td>
<td>United Health</td>
<td>262</td>
</tr>
<tr>
<td>DuPont</td>
<td>365</td>
<td>DuPont</td>
<td>247</td>
<td>Coca-Cola</td>
<td>243</td>
<td>Coca-Cola</td>
<td>261</td>
</tr>
<tr>
<td>McDonalds</td>
<td>305</td>
<td>McDonalds</td>
<td>220</td>
<td>JPMorgan Chase</td>
<td>206</td>
<td>JPMorgan Chase</td>
<td>175</td>
</tr>
<tr>
<td>Goldman Sachs</td>
<td>264</td>
<td>Goldman Sachs</td>
<td>169</td>
<td>DuPont</td>
<td>157</td>
<td>DuPont</td>
<td>157</td>
</tr>
</tbody>
</table>

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 4 has to realize that the results obtained by applying My Sentiment Api and the TextBlob are equal. The fifteen most popular companies evaluated for the first two tools been the same, varying only the order. The same happened to the SentiStrength and Affin tools also achieved the same type of evaluation. Then in Table 5 we present the returns obtained for the four of sentiment analysis tools. As expected returns of the first two tools (My Sentiment Api and TextBlob) are equal and the same with the last two tools (SentiStrength and Affin). In this test, we divided the budget into equal parts that is the initial investment in each company was equal.

<table>
<thead>
<tr>
<th>SENTIMENT ANALYSIS TOOLS</th>
<th>RETURN (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>My Sentiment API</td>
<td>0.88</td>
</tr>
<tr>
<td>TextBlob</td>
<td>0.88</td>
</tr>
<tr>
<td>SentiStrength</td>
<td>5.75</td>
</tr>
<tr>
<td>Affin</td>
<td>5.75</td>
</tr>
</tbody>
</table>

By analyzing the results of the previous table, we realized that the SentiStrength and Affin tools presented more positive returns than the others. With this case study can be
understood that the popularity of a company on social networks influence in some way the markets. Then another test will be performed within the calculated data. Table 6 presents the companies that have been in the top 15 for the four tools.

Table 6 - Six most popular companies in the four sentiment analysis tools.

<table>
<thead>
<tr>
<th>COMPANY</th>
<th>VOLUME OF TWEETS</th>
<th>COMPANY</th>
<th>VOLUME OF TWEETS</th>
<th>COMPANY</th>
<th>VOLUME OF TWEETS</th>
<th>COMPANY</th>
<th>VOLUME OF TWEETS</th>
</tr>
</thead>
<tbody>
<tr>
<td>MySentiment API</td>
<td></td>
<td>TextBlob</td>
<td></td>
<td>SentiStrength</td>
<td></td>
<td>Afinn</td>
<td></td>
</tr>
<tr>
<td>Microsoft</td>
<td>2863</td>
<td>Microsoft</td>
<td>1294</td>
<td>Microsoft</td>
<td>4918</td>
<td>Microsoft</td>
<td>4957</td>
</tr>
<tr>
<td>Visa</td>
<td>1054</td>
<td>Visa</td>
<td>563</td>
<td>Visa</td>
<td>698</td>
<td>Visa</td>
<td>827</td>
</tr>
<tr>
<td>Walmart</td>
<td>955</td>
<td>Walmart</td>
<td>712</td>
<td>Walmart</td>
<td>1024</td>
<td>Walmart</td>
<td>1098</td>
</tr>
<tr>
<td>Nike</td>
<td>964</td>
<td>Nike</td>
<td>587</td>
<td>Nike</td>
<td>376</td>
<td>Nike</td>
<td>412</td>
</tr>
<tr>
<td>Coca-Cola</td>
<td>730</td>
<td>Coca-Cola</td>
<td>422</td>
<td>Coca-Cola</td>
<td>243</td>
<td>Coca-Cola</td>
<td>265</td>
</tr>
</tbody>
</table>

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 7). 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. This last test achieved the best results as it has the four sentiment analysis tools to work together. The sentiment analysis tools to work separately do not have such good results because we have to count on a margin of error of analysis is not well done.

Table 7 - Return obtained by six most popular companies.

<table>
<thead>
<tr>
<th>RETURN (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.13</td>
</tr>
</tbody>
</table>

V. CONCLUSIONS AND FUTURE WORK

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 company’s 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 Afinn 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. 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.

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.
VI. REFERENCES


