Detecting Popularity and Innovation on Twitter to Find the Best stocks in SP500

João da Silva Franco
Instituto Superior Técnico
University of Lisbon
Lisbon, Portugal
joao.s.franco@tecnico.ulisboa.pt

Abstract— In the past years, social media has become one of the main sources related to the up and down movement of the financial market. The interest of investors in minimizing the risks when wagering in the financial market has increased with the possibility of predicting its up’s and down’s based on the sentiment classification of Twitter’s user’s opinions. In this paper, it is demonstrated how Twitter can be used to detect innovative and popular companies in the S&P500 index market. The collection of data is done through a robot that collects tweets from Twitter search API based on words, previously selected and related to those two subjects of interest. In order to analyse the text content of tweets and to classify a company’s innovation/popularity opinions either as a positive, negative or neutral opinions, it is implemented a Naïve Bayes classifier, with an accuracy of 70.6%. To infer which of the selected words are more influent in the detection of innovative and popular companies, a Genetic Algorithm is implemented. A forecasting of real-time outcomes of the S&P500 stock market values is calculated based on results obtained from the Genetic Algorithm, to detect which are the most innovative and popular companies for investors to wager. The best results were interesting, obtaining a profit of 11.8% for the Innovation case and 5.5% for the Popular case.

Keywords- Twitter, Stock Market, Genetic Algorithms, Naïve Bayes, Popularity/Innovation detection;

I. INTRODUCTION

Twitter is an online micro-blogging platform that enables users to post short messages with images and videos attached. This social network was launched in 2006 with more than 300 million monthly active users worldwide, generating over 500 million tweets per day. Twitter is usually used to share opinions and/or to discuss a variety of topics. The growth of relevant people in Twitter from different fields has increased since its launch and that is what makes this platform such an important tool to be studied. Consider the example of the U.S. 2008 elections where [1] focusing on people’s opinions through tweets, tried to predict who would win. Authors were right predicting that Obama would win. Analysts showed that Obama's initiative, in being the first presidential candidate to use social media as a major campaign strategy when Twitter was not being used by important personalities like it is today, has had a powerful impact in his election. Some cases, such as the previously mentioned, made researchers focus their attention on Twitter, aiming to use it as a predictive tool in different areas, through the analysis of people’s opinions. Important researches have been made in many different backgrounds such as in health care sector, sports, political aspects and financial market.

Some researchers thought about Twitter as a social web platform that could give a new perspective to market analysts. They thought that through the analysis of Twitter’s user’s opinion it would be possible to gather information related to the stock market, and use it to predict stock market movements. Due to the risk and complexity that involves investing in the financial market several techniques are being developed to minimize investor’s losses and increase their profits. The concept behind these techniques is to use machine learning algorithms to correlate people’s opinions about the financial market with the stock market movements, aiming to predict possible future changes on it.

This work will focus on the impact that Twitter has on the financial market, using it to predict the up and down movements of the S&P500 Index. Focusing on tweets information, a genetic algorithm is implemented in order to detect the evolution of the S&P500 Index, which analyses the possibility of using the popularity and innovation on Twitter as an indicator to predict financial stock market movements and to get profits.

The main contributions of this work the creation of a dictionary of popular and innovative words to filter tweets in order to obtain valuable data and also a definition of investment strategies to correlate the data with the stock market movements, generating profit.

In this paper Section II presents the state-of-the-art for the concepts of Twitter, Sentiment Analysis, Genetic Algorithms and Naïve Bayes. Section III describes the proposed system’s architecture. In Section IV the case studies and results are discussed. Conclusions and Future Work are addressed in Section V.

II. RELATED WORK

A. Twitter

Twitter is a social web platform that allows registered users to follow and communicate with other users through short messages. This platform allows one to follow other’s account and, at the same time, those who have been followed can decide to follow back or not. Each user’s profile contains the
sentiment analysis can have in a company that seeks to improve its marketing strategy by identifying their client’s opinion concerning their brand. Amongst Twitter or marketing there are other different areas where sentiment analysis can play an important role such as in health care sector, sports or in political aspects.

1) Sentiment Analysis on Twitter

The interest on analysing people’s opinion through tweets content increased in the last years as a result of some first and important researches made in this field [4] [5]. In [4], authors report that their research was in benefit of consumers who want to search for opinions of others about products before buying, or companies that want to know their client’s opinion as concerns the brand. The authors focused on the sentiment analysis of tweets by developing a method in order to classify tweets as either positive or negative. Since there are a variety of topics discussed on Twitter it would be difficult and expensive to build a model able to classify the polarity of each tweet, thus, the idea was to collect data based on tweets with emoticons in order to detect the general sentiment expressed. Using Twitter APIs, to extract tweets with emoticons in it, the authors divided their training data in two classes: positive tweets with positive emoticons “:)” and negative tweets with negative emoticons “:(”. Following the work developed by [6], the machine learning techniques used for the analysis were Naïve Bayes, Maximum Entropy, and Support Vector Machines, concluding that the last outperformed the other classifiers. The feature extractors tested were Unigram, and Bigram in combination with Unigram and parts-of-speech (POS) features. The authors stated that the use of unigrams as features outperforms both models and, being consistent with [6], they noticed that the use of POS tags as features was not useful. In [5], authors continued the study of sentiment analysis in Twitter and followed the same procedure as in [7] and [4] by classifying tweets of user’s accounts of newspapers and magazines companies of North America whether positive, negative, and neutral. Finally, to detect the polarity of tweets they tested the use of three algorithms: SVM, Naïve Bayes and CRF. They report that the accurate method for this analysis is to use the multinomial Naïve Bayes classifier that uses N-gram and POS-tags as features.

2) Influence on Twitter

We live in a world where the majority of people are influenced by what they see, read and hear. “Following the other” is currently a tendency, and the same happens on Twitter, when one posts a tweet it will be read and retweeted by the followers. Hence, for instance, if one wants to predict something, one has to know which parameters influence that cause. To better understand or predict the stock market changes it is necessary to identify who has a bigger influence on it, including characters or companies.

In [8], authors have investigated how influential a Twitter’s user can be based on the content that one posts. The authors analysed 12 different users, including celebrities, news outlets and social media analysts. This research brought some interesting conclusions: celebrities with a higher number of followers promote more conversation than retweetable content and that News outlets are more influential to other users than

---

correspondent biography, photo, website and location, and those are public for everyone, including Twitter’s non-members. Twitter allows its users to share additional information such as the date of birth, followers and following people. Tweets can go up to 140 characters. Since tweets can be sent in real-time to followers, they might be classified as instant messages. The difference between one and another remains in the fact that tweets are posted on Twitter’s website and, that makes tweets a permanent, searchable and public information to anyone, whether they are a Twitter member or not. A Retweet (RT) is a re-posting of a tweet. This feature is used to share someone’s tweet with followers and it can be done with own tweets or one following user’s tweets. Hashtags (#), which can be included in tweets are used to denote a topic of conversation and can be very helpful to find tweets based on topics. Mentions (@) are a way to reference a user by his/her username. Twitter also provides the feature Cashhtag ($), which is used to identify a company’s stock symbol. Consider as an example the Google’s Cashhtag: it is the Google’s stock symbol preceded by the dollar sign ($GOOGL).

1) Twitter as an outcome predictor for the financial market

When investing there are always high risks either when wagering or exiting the stock market. The use of computational methods as a tool to help investors on how and when to invest in the stock market have increased in the past years. A relevant research made by [2] investigated whether measurements of collective mood states derived from large-scale Twitter feeds are correlated to the value of the Dow Jones Industrial Average (DJIA) over time. They analysed the content of tweets by two mood tracking tools. OpinionFinder to measure positive vs negative mood and Google Profile of Mood States to measure mood in terms of 6 dimensions (Calm, Alert, Sure, Vital, Kind, and Happy). The research shown that changes in the public mood state can indeed be tracked from the content of large-scale Twitter feeds and help to predict the value of DJIA in certain events. Following the previous work done in [2], authors in [3] 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. Authors in [3] proposed a new cross validation method for financial data, obtaining 75.56% accuracy using Self Organizing Fuzzy Neural Networks on the Twitter feeds and DJIA values. They also developed a strategy, using a Naïve Bayes algorithm, to maintain a profitable portfolio.

B. Sentiment Analysis

Sentiment analysis is the computational process of identifying and classifying sentiments expressed in a text, in order to determine the writer’s opinion towards a certain topic. Hence, for instance, one who wants to acquire a laptop will search in web forums aiming for online opinions or reviews of different brands and models. After evaluating the content available, one has to decide which laptop to buy based on the overall sentiment expressed. Sentiment analysis is used for the same purpose. Actively use technological information, such as tweets, to investigate and understand people’s opinion in many different areas and topics. This demonstrates the massive impact that sentiment analysis can have in a company that seeks to improve its marketing strategy by identifying their client’s opinion concerning their brand. Amongst Twitter or marketing there are other different areas where sentiment analysis can play an important role such as in health care sector, sports or in political aspects.

---

- **Public** for everyone, including Twitter’s non-members.
- **Retweet (RT)** is a re-posting of a tweet.
- **Hashtags (#)** can be included in tweets.
- **Mentions (@)** reference a user by their username.
- **Cashhtag ($)** identifies a company’s stock symbol.
- **OpinionFinder** measures mood.
- **Google Profile of Mood States** measures mood in 6 dimensions.
- **Sentiment Analysis** helps predict stock market values.

---

Additional notes:
- **Twitter** shares additional information like date of birth, followers, and location.
- **Hashtags (#)** denote a topic.
- **Mentions (@)** reference users.
- **Cashhtag ($)** identifies stock symbols.
celebrities considering the large amount of retweetable content published. Authors also analysed the influence of a user based on the ratio between the number of followers and the number of followees. They concluded that this parameter cannot give an accurate measure of influence on Twitter but, can be useful to separate Twitter’s community into different types of user. As stated before, Twitter has been identified as a major influential factor to the financial market movements.

Authors in [9] developed a method to identify relevant Twitter users aiming to form a financial community aside of sentiment analysis to provide a better prediction of financial market movements. They stated that financial market movement is a manifestation of the market and behaviours toward future outcomes. Authors also analysed the influence of a user based on the ratio between the number of followers and the number of followees. They concluded that this parameter cannot give an accurate measure of influence on Twitter but, can be useful to separate Twitter’s community into different types of user.

C. Genetic Algorithms

Genetic Algorithms are a process of natural selection. This process starts with the selection of the fittest individuals from the population. They produce offspring which acquire the characteristics of the parents and will be added to the next generation. If parents have better fitness, their offspring will be better than parents and have a better chance at surviving. This process is repeated until the generation with the fittest individuals is found. Five phases are considered in a Genetic Algorithm:

- **Initial generation**: set of individuals which is called a population, where each individual is considered a solution for the problem under study, characterized by a set of parameters known as genes that jointed form a chromosome (solution).

- **Fitness Function**: determines how fit an individual is represented by the ability of an individual to compete with other individuals. It gives a fitness score to each individual of the population. This fitness score represents the probability of an individual to be selected for reproduction.

- **Crossover**: used to generate two new chromosomes from a pair of chromosomes. The new chromosomes are produced by the combination of genes of their parents, see Figure II.1.

- **Mutation**: defined as a small modification applied to a chromosome’s feature aiming to obtain new possible solutions. This operation is used to generate chromosomes with different characteristics, introducing diversity in the genetic population and is usually applied with a low probability, see Figure II.2.

- **Survivor selection**: define which chromosomes are going to be used to generate the next generation of chromosomes or which are going to be discarded. This generation process is repeated until the Genetic Algorithm reaches the conditions towards the optimal solutions.

  1) **Genetic Algorithms to Forecast Financial Markets**

These algorithms have been used for many different purposes and in many different areas such as the financial market sector [10] [11]. The goal of this algorithm is to achieve the best solution within a considerable number of possible solutions.

In [10], authors proposed the use of a genetic algorithm to detect which features better represent an influential user on Twitter towards the rise and fall of the shares’ value. A sentiment classifier was implemented to analyse the collected tweets for this research. Final results show that the most influential Twitter users are those that have a higher number of tweets and having an increasing number of new followers. Also features such as the number of tweets and the ratio between followers and friends were characterized as important for the final result obtained. In [11], authors developed a system to predict stock returns of the Dow Jones Industrial Average based on tweets posted by a financial community. A genetic algorithm was implemented in order to get the best solution for the portfolio optimization problem, aiming to select a set of stocks that could give a positive return with a reduced risk. Authors state that the tweets collected using the financial community were very important when detecting important events in the life of companies. They achieved an 8.13% profit based only on the volume of positive tweets collected for each company.

D. Naïve Bayes Classifier

In this section we introduce the Naïve Bayes classifier. This classifier is frequently used to classify tweets in a sentimental way, detecting whether tweet’s content are transmitting a positive, negative or neutral opinion about the subject of matter. These classifiers study the classification task from a statistical point of view, meaning that the probability of a class C is given by the posterior probability $P(C|D)$ given a training document D. The document D corresponds to all of the text in the entire training set. It is given by $D = (d_1, d_2, ..., d_n)$, where $d_i$ is the $i^{th}$ word of document D. Using Bayes’ rule, the posterior probability can be rewritten as:

$$P(C = c_i|D) = \frac{P(D|C = c_i) \cdot P(C = c_i)}{P(D)}$$

(1)

Since the marginal probability $P(D)$ is equal for all classes, it can be disregarded and the equation becomes:

$$P(C = c_i|D) = P(D|C = c_i) \cdot P(C = c_i)$$

(2)

The document D belongs to the class C which maximizes this probability, so:

$$C_{NB} = \text{arg} \max P(D|C) \cdot P(C)$$

(3)

$$C_{NB} = \text{arg} \max P(d_1, d_2, ..., d_n|C) \cdot P(C)$$

(4)

Assuming conditional independence of the words $d_i$, this equation simplifies to:
\( C_{NB} = \text{argmax } P(C) \cdot \prod P(d_i|C) \)  

(5)

The probability \( P(d_i|C) \) is the conditional probability that the \( i^{th} \) word belongs to class \( C \). This probability corresponds to the frequency of word \( i \) in class \( C \) relative to the total number of words in class \( C \).

\[
P(d_i|C) = \frac{\text{count}(d_i, C) + 1}{\sum (\text{count}(d_i, C) + 1)}
\]

(6)

This machine learning algorithm let us estimate the prior-probabilities with a training set of documents that are already labelled with their classes. With this training set we can train the model and obtain values for the prior probabilities. This train model can then be used to classify the collected tweets for this research.

### III. SYSTEM ARCHITECTURE

The aim of this module is to identify which words have more influence when characterizing the innovation and popularity of a company included in the S&P500 index, based on a pre-selected word dictionaries defined for the subject innovation and popularity. The principle behind this study is to understand the impact of a tweet that relates to a company’s popularity or innovation using any of the pre-selected words.

Therefore, it is intended to develop a system able to detect which words better characterize a company’s popularity and innovation, for this, positive tweets that contain a specific innovative/popular word will lead to an increase on the stock market of that company, while negative tweets will result on a decrease of the stocks. To build this system it is necessary to collect tweets considered as innovative or popular using the pre-select dictionaries and analyse it with the help of a Genetic Algorithm to detect which words have more impact on the stocks of the companies under study.

#### A. General Architecture

This work is divided into different modules designed to complete each task and to fulfil the purpose of this system. These modules combine and communicate with each other by exchanging data throughout the system process. The combination of the modules considered for this process are illustrated in Figure III.1 and represent the architecture of the system used to accomplish the proposed goal.

The Data Sources module is used to communicate with external sources as:

- Twitter: used to get tweets’ information related to companies under study.
- Yahoo! Finance: online tool owned by Yahoo used to gather information about stocks quotes of companies in the S&P500 index.
- Google Finance: Web-based application owned by Google used to collect stock information about any company of interest.

The Data Collection module is where the data is collected from the external sources identified in the Data Sources module, and stored for future analysis in the Data Analysis module. The Data Analysis module processes the information prepared in the Data Collection module to be considered by the algorithms. Afterwards, this will be used in the Portfolio Management module as the basis to develop the investment strategy to be adopted in this system based on long and short strategies. The Output Data module displays the results obtained for each investment strategy considered, comparing its outcome with the S&P500 index and highlighting the words that better characterize the innovation and popularity of a company in the S&P500 index for each period of analysis.

![Layered architecture of the system.](image)

#### 1) Twitter Data Collection

This module is responsible for the collection and creation of the tweets’ databases to consider in this study. The most common tool for this task is the Twitter search API however, their free tier limits the user to search for tweets published in the last 7 days. The use of this platform would implicit a weekly collection of data for a long period to obtain the desired amount of information. There are however, paid tools that provide access to older tweets.

In order to surpass this issue by accelerating the process of collection, it is used a robot written in python, named “Get Old Tweets” [12], that bypasses some limitations of Twitter Official API. The mechanism used by this robot is based on the input defined, meaning that if the company in search is Apple the robot searches tweets with the cashtag symbol ‘$AAPL’ and saves all related tweets by scrolling down through Twitter page. Twitter also provides users the chance to search tweets written in English and to specify the period of collection by selecting the dates. All these features can be used by the robot to capture the information required.

To obtain some consistency in the subject of search it is necessary to define a dictionary of words that relate to innovation and popularity. To better understand this section, consider that the word “cool” which is selected for the dictionary of popularity. During a certain period tweets of a specific company containing the cashtag of that company (e.g.: $AAPL, Apple’s cashtag) and the word “cool” are collected. Based on the analysis of this content it will be possible to understand the impact of that word in, for example, Apple’s popularity. If there are 100 tweets under these circumstances and the majority of them positively reflecting Apple, its stocks...
increase meaning that “cool” is a strong word to define a company’s popularity in the market. Thus, it is always required to compare this information with the financial data collected.

Based on this information, it is not possible to predict which of the selected words will turn to be strong words. To ensure some consistency both dictionaries were created based on a few base words and their synonyms avoiding with this, a big number of non-related words between them. Table III.1 shows the popular and innovative dictionary produced based on internet synonyms webpages and available dictionaries, each one with a total of 32 words.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Dictionary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Popularity</td>
<td>popularity; cool; favourite; best; trending; trendy; successful; favorite; trend; amazing; reputation; popular; trendsetter; attractive; hot; success; trendiness; profit; leader; competitive; evolution; influential; lucrative; coolest; notorious; quality; reliable; good; likeable; profitable; prestigious; liked</td>
</tr>
<tr>
<td>Innovation</td>
<td>innovation; new; creative; pioneer; inventors; unusual; different; novelty; release; launch; innovatory; innovators; novel; innovational; modern; brand-new; newness; inventive; futuristic; latest; fresh; precedent; innovating; original; idea; developed; innovative; originative; pioneering; visionary; innovator; unique</td>
</tr>
</tbody>
</table>

2) Financial Data Collection
This module represents the collection of all the information regarding the company’s stock quotes in order to understand the impact of opinion mining in the stock values of a company. This module is of an extremely importance in the detection of which words better characterize an innovative and popular company. The financial data is collected using a python library for the Yahoo Finance API, named “yahoo finance” [13], and for the Google Finance API, named “googlefinance” [14].

For this study the variable used to measure the impact of people’s opinion in the stock market is the OpenPrice, which represents the price of a company’s stock at the beginning of the trading day. Based on this value, the stock price variation, in percentage, of a company C between date d1 and d2 is calculated using the formula (7).

\[
\% \text{ Gain}_{\text{OpenPrice}}(C, d_1, d_2) = \frac{\text{OpenPrice}(C, d_1) - \text{OpenPrice}(C, d_2)}{\text{OpenPrice}(C, d_2)}
\] (7)

3) Sentiment Analysis Processing
The purpose of this module is to develop a classifier able to analyze and classify people’s reaction when posting about the innovation/popularity of a company in the S&P500. To achieve an accurate outcome it is necessary to prepare and select the tweets for the classifier as many of the tweets collected from the API might not be ‘classifiable’ into some sentiment.

This module is composed in two phases: Tweet cleaning and Sentiment Analysis.

a) Tweet cleaning
The first phase has the challenge of taming the data by removing irrelevant expressions used by the users that can lead to an inaccurate sentiment classification. This tweet cleaning task consists of 7 subtasks to accomplish this process: Removing emoticons and additional punctuation marks; Removing URL’s and pictures; Converting all tweets into lower case and lengthening; Removing hashtags and cashtags; Removing mentions; Removing stopwords; Removing special characters. The order of each subtask must be prioritized in order to avoid losing important information when cleaning the data for analysis.

b) Sentiment Analysis
The second phase consists on training the classifier based on pre-classified tweets. This second phase is crucial in detecting which words better characterize an innovative and popular company by outputting the sentiment classification of tweets that contain any of the pre-selected words included in the innovative and popular dictionary. The first step of the sentiment classification is training the Naive Bayes classifier. In order to proceed it is necessary to define a dataset so the classifier can learn from, as well as a test dataset to ensure that the accuracy of the classifier meets the expected standards. To proceed with the training it is required to select a database of labelled tweets with positive and negative sentiment polarity.

To define and collect the database for the training process it was used Twitter hashtags as labels to the emoticon expressed by tweets. For example, the hashtag “#happy” is labeled as a positive tweet as it transmits a happy and positive feeling. The same mechanism was used to build the negative dataset of tweets. To create these two databases, mostly based on the “Model of Basic Emotions” used in [15], it was defined a list of 80 hashtags, corresponding to each class of emotion with 40 positive and 40 negative hashtags. All this information was collected in the period between 15th December 2014 and 20th December 2016 using the Get Old Tweets robot. Then, tweets were submitted through a data clean process, previously explained, remaining only useful information/words for the sentiment classification. The output of this analysis is a list of words ordered by its frequency of appearance in tweets under study. The Naive Bayes classifier uses this list to calculate the sentiment polarity of each tweet based on the frequency of appearance assigned to each word for a positive or negative class.

Based on the list developed during the training process, with all the prior probabilities and likelihoods that correspond to the frequency of a word to be referred in a positive or negative class, the classifier is going to sum all the positive and negative probabilities of each word, represented in tweets under study, and output the final positive and negative probability using functions explained in section II.D. Once all these calculations are computed the Naive Bayes classifier will select the category with the higher prediction rate as the result of our classification. Every “new word” that does not exist in the training set will have a probability as zero since they were not included in the classifier’s dictionary.
Due to the high amount of time that the classifier takes to process the sentiment polarity of each tweet it is necessary to reduce or understand if, by removing words with less impact on the final sentiment polarity, it is possible to improve the classifier’s accuracy and reduce the amount of time to output the final result. To test the accuracy associated to each dictionary of words it was used a dataset called “Twitter Sentiment Analysis Dataset” containing 7086 sentences, with 3995 sentences labeled with positive sentiment and 3091 with negative sentiment [16]. The best accuracy obtained was 70.6% using only 5% of words with the highest frequency to appear either in a positive or negative tweet.

To avoid buying stocks based on tweets classified as neutral opinions it was necessary to establish a difference between a positive and a negative opinion. To achieve this objective it was included the Neutral class. Based on [17] and due to the necessity of establishing a boundary value to classify tweets with positive, neutral or negative sentiment, each tweet with sentiment polarity under the value 0.22 and above -0.22 is classified as a neutral opinion. The summary of the classifier decision values are presented in formula (8).

\[
\text{Classification} = \begin{cases} 
  x < -0.22, & \text{Negative} \\
  -0.22 < x < 0.22, & \text{Neutral} \\
  x > 0.22, & \text{Positive} 
\end{cases}
\]  

\[(8)\]

4) Word Features

This module is essential to define the variables that will later provide information in order to identify which words better describe a popular and innovative company in the S&P500 index. The variables defined in this section will help the Genetic Algorithm to output the dictionary of popularity and innovation words that better describe a company in those two subjects.

To better understand this module it is important to assume that each word was used to collect a considerable number of tweets. This was the approach considered for this study in order to have a consistent database of tweets related to each word, to obtain a final value of influence for each word. For example, to understand the influence of the word “popular” in companies, tweets were collected during a specific time period of analysis with text information related to the hashtag of the companies under study containing the word “popular”. This was the strategy defined to understand the impact of the word “popular” in companies under study. The same process gets repeated for all the words included in the popular and innovative dictionaries. The following features were selected based on which variables might have impact on the two subjects under analysis:

- **Word Activity**: This feature is important to understand how often the word i is used to describe the company C in terms of popularity and innovation, see formula (9).

\[
\text{WordActivity}_{i,C} = \frac{\# \text{tweets about company } C \text{ with word } i}{\# \text{tweets of company } C} 
\]  

\[(9)\]

- **Hashtags per Word**: This feature represents how often people use hashtags in tweets containing the word i to describe the company C in terms of popularity and innovation, see formula (10).

\[
\text{HashtagWord}_{i,C} = \frac{\# \text{hashtags in tweets of company } C \text{ with word } i}{\# \text{tweets of company } C} 
\]  

\[(10)\]

- **Likes**: This feature represents an average of how much people “liked” tweets containing the word i used to describe the company C in terms of popularity and innovation, see formula (11).

\[
\text{Likes}_{i,C} = \frac{\# \text{likes in tweets of company } C \text{ with word } i}{\# \text{tweets of company } C} 
\]  

\[(11)\]

- **Mentions by Word**: This feature represents the average number of mentions done by users in tweets containing the word i when describing the company C in terms of popularity and innovation, see formula (12).

\[
\text{Mentionword}_{i,C} = \frac{\# \text{mentions in tweets of company } C \text{ with word } i}{\# \text{tweets of company } C} 
\]  

\[(12)\]

- **Words per Tweet**: This feature represents the average number of words used in tweets containing the word i when describing the company C in terms of popularity and innovation, see formula (13).

\[
\text{WordsTweet}_{i,C} = \frac{\# \text{words in tweets of company } C \text{ with word } i}{\# \text{tweets of company } C} 
\]  

\[(13)\]

- **Retweets by Word**: This feature represents how many retweets have been done in tweets containing the word i when describing the company C in terms of popularity and innovation, see formula (14).

\[
\text{RetWord}_{i,C} = \frac{\# \text{retweeted tweets of company } C \text{ with word } i}{\# \text{tweets of company } C} 
\]  

\[(14)\]

- **Different Users that Retweeted**: This feature represents the number of unique users that have retweeted tweets containing the word i used to describe the company C in terms of popularity and innovation, see formula (15).

\[
\text{D.U.R}_{i,C} = \frac{\# \text{unique user retweeted tweets of company } C \text{ with word } i}{\# \text{tweets of company } C} 
\]  

\[(15)\]

5) Genetic Algorithm

The aim of this module is to discover what characterizes the words that better describe a popular and innovative company in the S&P500 index based on the pre-selected dictionaries. Based on the features explained in the previous section the Genetic Algorithm (GA) will output the ideal values associated to each feature that represent the “influential words”.

This process begins with an asset of individuals which is called a population. Each individual is a solution to the problem under study. An individual is characterized by all the 7 features, known as genes. Genes are joined into a string to form a chromosome (solution) and is illustrated in Figure III.2, in which each gene is a word’s feature.

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Gene 1</th>
<th>Gene 2</th>
<th>Gene 3</th>
<th>Gene 4</th>
<th>Gene 5</th>
<th>Gene 6</th>
<th>Gene 7</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Word Activity</td>
<td>Hashtags per Word</td>
<td>Likes</td>
<td>Mentions by Word</td>
<td>Words per Tweet</td>
<td>Retweets by Word</td>
<td>Different Users that Retweeted</td>
</tr>
</tbody>
</table>

Figure III.2 - Representation of a chromosome used within this study.

The number of individuals is represented by the sum of the number of words assigned to each company under study. Since each word will be assigned to different values in each feature.
the aim is to obtain the best range of values for each feature. To perform this task it is necessary to obtain sets of words for each feature where each set is represented by a certain number of decreasing subsets, called ranks. For example, the feature “Word Activity” set is sorted so the first ranked subset contains the words with more related tweets posted and the last subset with the words with less activity. This process is then applied to all features. This will be useful so the GA can select within each feature the best range of values that determine the words with more influence in the stock market. The number of subsets will be calculated by the GA, testing a different number of possibilities. In Figure III.3 it is presented a possible example of how this process will happen. Based on n ranks the GA will output the optimal word by mixing all the genes of the words under study. The result will be the best rank for each feature, representing the best optimal range of values for each feature.

\[
\text{Fitness}_{\text{individual}} = \sum_{\text{word}=1}^{n} \sum_{\text{day}=1}^{m} \left( \% \text{Gain}_{\text{openprice}}(C; \text{day}_{\text{after}}; \text{day}) \times \sum_{\text{tweet}=1}^{p} \text{polarity}_{\text{tweet}} \right)
\]

(16)

Figure III.3 – Representation of the sets of features

- **Initial Generation:** To create the initial generation it is necessary to set the number of ranks chosen by the system. This value must be between 10 and 30. It should not be less than 10 in order to obtain a different solution with a range of values associated to different words reducing the probability to have few subsets containing almost the entire sample and not more than 30 to avoid a big dispersion of values. After defining the number of ranks each feature will be divided in to, it is required to randomly initialize the population by assigning randomly ranks to all individuals.

- **Evaluation Function:** In order to calculate the fitness function the program will take into consideration the historical data collected about the stock market. Since only the opening price is considered for this study it is crucial to take into account the timestamp when each tweet was posted. In this way, only tweets posted between 2:30 p.m. the previous working day and 2:30 p.m. the current day are considered for the analysis of current’s day opening price. The reason for this is that trading hours for S&P500 stocks starts at 9:30 a.m. UTC-05:00 and closes at 4:00 p.m. UTC-05:00, excluding extended hours trading. The evaluation is then calculated through the formula (16) and measures the impact on the stock market of words assigned to a specific rank.

The function \(\% \text{Gain}_{\text{openprice}}(C; \text{day}_{\text{after}}; \text{day})\) gives the stock value variance of the company \(C\) between the day \(d\) of the chosen sliding window and the day \(d_{\text{after}}\) for which the system wants to check the impact of a word. The inner summation gives the word sentiment coefficient about company \(C\) of all tweets \(p\) for each day \(m\) of the sliding window. This value is then multiplied by the function \(\% \text{Gain}_{\text{openprice}}\) for each day of the sliding window, representing the variation of the stock price of company \(C\) between the day \(d\) and \(d + d_{\text{after}}\) and obtaining the score of that word for that day \(d\). The score is then calculated for each day of the sliding window defined and for all the words of each rank, obtaining the fitness score for each rank.

- **Selection:** To perform this task it is employed the method called “Tournament Selection” which consists in randomly picking \(k\) individuals from the population, the fittest of them (winner of each tournament) is selected for crossover and able to reproduce. The remaining individuals are returned to the population to be picked up again. The same process happens to obtain the other parent of reproduction. To avoid losing diversity in the population the tournament size selected for the GA is 5, which is a small value compared to the size of the population.

- **Crossover:** To perform this task it is employed the method called Double Point Crossover, which consists in picking randomly two crossover points from the parent chromosomes. The bits in between the two points selected are swapped between the parent chromosomes.

- **Mutation:** It was used a mutation rate of \(1/n\), where \(n\) is given by the number of genes. This is a frequently used mutation rate that means, on average, one mutation will occur in every offspring.

6) **Investment Strategies**

In order to improve our investments it was necessary to define some rules of investment. In order to obtain a consistent portfolio of investment it was defined two long strategies and two short strategies. Since there is no budget limitation to buy stocks, the system will purchase stocks from different companies in the same period of time and every time it is detected something positive in terms of popularity or innovation. The only limitation is that it is only allowed to have one stock per company.

a) **Long Strategies**

Long equity is an investment strategy that takes long positions in stocks that are expected to rise. The two long strategies defined for this work are given following:

- **Top-10 Companies:** This strategy is based on people’s opinion on Twitter related to a company’s popularity/innovation. For each day of the testing set it is computed the sentiment coefficient of each company under study and are bought stocks of the Top-10 companies.

- **Long Influential Words:** For this strategy, it is computed every day’s average sentiment coefficient of each company during the testing period. If this average is greater than the minimum value established for an opinion to be classified
as positive (sentiment coefficient > 0.22), then the stock of that company is purchased.

b) Short Strategies

Short equity is an investment strategy that takes short positions in stocks that are expected to decline. The two short strategies defined for this work are given following:

- **Short after certain days**: The aim of this strategy is to sell a stock’s company, previously bought, after a certain number of days. This approach does not represent public’s opinion on Twitter but an alternative to that, since it is intended to test the validity of Twitter’s popular/innovative opinions about companies under study. In this strategy, the sale of stocks is tested after 2, 5 and 10 working days after the stock’s acquisition.

- **Short Influential Words**: This sale strategy is similar to the Long Influential Words strategy where each company is analysed separately. For each day of the testing period it is executed an average sentiment coefficient of each company and if that value is lower than the minimum value established for an opinion to be considered as a negative (sentiment coefficient < -0.22), then the stock of that company is sold.

## IV. SYSTEM VALIDATION

### A. Implementation Parameters

It is necessary to define a set of parameters when analysing which words better describe S&P500 companies in terms of popularity and innovation. In this simulation it is only possible to have one stock per company at the same time. When the portfolio purchases a stock it has a minimum holding period of one day, taking into consideration that in this work are only tested investments at the opening times of the market. For example, if the portfolio purchases one stock from Apple it will have a minimum time of acquisition of one day, selling it in the next day open time. The number of days defined to measure the impact of a word in this simulation is 7 working days. In [18] and [19] authors tested 10 and 5 days respectively to measure the impact of users’ opinions on Twitter. Since the results were not as good as expected in both studies the approach in this simulation is 7 days period, starting in the day the tweet is posted until 7 working days after.

#### a) Evaluation Metric

To evaluate the performance associated to each popular/innovative word and to understand the impact of using it to classify a company’s popularity or innovation it was used the Return of Investment (ROI) metric. This function outputs, in percentage, the profit of an investment. In this study, it was used the ROI of S&P500 index as benchmark to understand which words are more influential when characterizing a company’s popularity/innovation, and 2 weeks of testing with investments strategies based on the outcome obtained from previous 3 weeks of training. Figure IV.1 illustrates the approach used in this work.

![Figure IV.1 – Sliding Window approach for 2 simulation periods.](image)

**B. Case Study I**

The first case study is based on tweets collected between September 2016 and February 2017 as a training and testing set. This case study will be the benchmark of this work, since it uses all the words included in the popular and innovative dictionaries. For example, if a tweet of a certain company has a word included in the popular dictionary it will be automatically classified as a popular tweet of that company and, furthermore, used to invest in that company’s stock market based on that tweet’s sentiment polarity. The aim of this study is to detect which words better describe a popular or innovative company. Each word is linked to a value that refers to the price variation of a company’s stock’s market value. This variation represents the company’s stock value when the tweet is posted minus the stock value 7 days later, to observe the impact/influence of the use of that word on twitter in a company’s stock market value. This process is repeated for each word included in the popular/innovative dictionary and for all the companies included in the S&P500 index to form the basis of decisions that are taken.

This first case study is based in four different strategies of investment. There are two long strategies and two short strategies, those form the following methods to invest based on our collection of popular and innovative tweets. The best return of investment obtained considering this case study was a profit

\[
ROI(\text{Return of Investment}) = \frac{\text{Gain} - \text{Cost}}{\text{Cost}}
\]

Finally, the profit is given by calculating the average of all investment returns during a specific period, see formula (18).

\[
\text{Profit} (%) = \frac{\sum_{i=1}^{n} \text{ROI}_i(\%)}{n}
\]

To obtain a more accurate view of these investments and to understand the quality performance of any strategy of investment considered for this study it is necessary to compare its ROI with the ROI of S&P500 index. The ROI of S&P500 index simulates someone that buys all the companies present in the index and holds it during the selected period.

#### b) Analysed Companies

In order to obtain a more accurate result it was took into consideration companies with more than 100 tweets for each sliding window considered in this study. This number represents the average number of tweets related of all companies and the minimum number of tweets that one company needs to have in each sliding window so it can be a subject of analysis.

#### c) Training and Testing

As mentioned in previous chapters Twitter’s user tendency and influence are always changing. For example, considering a financial community of influential users to be the basis of this system investments it might be good but it is important to update it depending on the time period. The same happens with words, some might be good indicators of a company’s influence for a specific time period however, it is always important to validate it on a regular basis in order to obtain more accurate results when investing based on it. The system validation of this study is based on a sliding window process with 3 weeks of training, to understand which words are more influential when characterizing a company’s popularity/innovation, and 2 weeks of testing with investments strategies based on the outcome obtained from previous 3 weeks of training.
of 10.8% for the Innovation study and 3.5% for the Popular study, both considering the strategy “Top-10 Companies & Short after certain days” and selling stocks after 2 days its purchase. See Figure IV.2 and Figure IV.3 for Popular and Innovation case respectively.

C. Case Study II

The aim of this second case study is to improve the results previously obtained in case study I using a genetic algorithm. This algorithm presents some complexity and it is used to return the characteristics of the selected words from both dictionaries (popular and innovative) that were more influence on the stock market. This approach was applied to each sliding window, based on data collected between September 2016 and February 2017. In this case study it was considered the use of 10 ranks to divide the population into subsets of words. To identify which words are more influential in detecting popularity and innovation on Twitter to find the best stocks in the S&P500 index market it was decided that an influential word must have, at least, 3 feature values within the 7 total optimal features ranges outputted by the GA. The perfect analysis would be to consider only words containing features values between those ranges however, it is not possible due to the lack of information available, since there are not many words with those characteristics. For example, for each testing sliding window only words with, at least, 3 features values within the optimal features ranges outputted by the GA will be considered. The best return of investment obtained considering this case study was a profit of 11.8% for the Innovation study and 5.5% for the Popular study, both considering the strategy “Top-10 Companies & Short after certain days” and selling stocks after 2 days its purchase. See Figure IV.4 and Figure IV.5 for Popular and Innovation case respectively.

Table IV.1 - ROI values obtained for each investment strategy used in the two case studies.

<table>
<thead>
<tr>
<th>Investment Strategies</th>
<th>Case Study I</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Case Study II</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Top-10 Companies &amp; Short after certain days</td>
<td>2 days</td>
<td>5 days</td>
<td>10 days</td>
<td>2 days</td>
<td>5 days</td>
<td>10 days</td>
<td>2 days</td>
<td>5 days</td>
<td>10 days</td>
<td>2 days</td>
</tr>
<tr>
<td></td>
<td>10.8%</td>
<td>4.8%</td>
<td>4.2%</td>
<td>3.5%</td>
<td>3.2%</td>
<td>3.3%</td>
<td>3.5%</td>
<td>3.2%</td>
<td>3.3%</td>
<td>3.3%</td>
</tr>
<tr>
<td>Top-10 Companies &amp; Short Influent Words</td>
<td>4.6%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Long Influent Words &amp; Short after certain days</td>
<td>2 days</td>
<td>5 days</td>
<td>10 days</td>
<td>2 days</td>
<td>5 days</td>
<td>10 days</td>
<td>2.7%</td>
<td>9%</td>
<td>4.4%</td>
<td>2.4%</td>
</tr>
<tr>
<td></td>
<td>5.2%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Long Influent Words &amp; Short Influent Words</td>
<td>5.2%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Investment Strategies</td>
<td>Case Study II</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Case Study II</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top-10 Companies &amp; Short after certain days</td>
<td>2 days</td>
<td>5 days</td>
<td>10 days</td>
<td>2 days</td>
<td>5 days</td>
<td>10 days</td>
<td>11.8%</td>
<td>5.8%</td>
<td>5.2%</td>
<td>5.5%</td>
</tr>
<tr>
<td>Top-10 Companies &amp; Short Influent Words</td>
<td>6%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Long Influent Words &amp; Short after certain days</td>
<td>2 days</td>
<td>5 days</td>
<td>10 days</td>
<td>2 days</td>
<td>5 days</td>
<td>10 days</td>
<td>3.6%</td>
<td>9%</td>
<td>5.8%</td>
<td>2.4%</td>
</tr>
<tr>
<td></td>
<td>7.4%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table IV.1 sums up all the percentage values of ROIs obtained for each investment strategy used for both case studies considered in this study. It is possible to observe that the best ROIs were obtained in the second case scenario. This is due to the capacity of Genetic Algorithms to analyse only tweets with relevant information about company’s popularity and

![Figure IV.2 - Popularity: Daily ROI for S&P500 index and Profit for strategy ‘Top-10 Companies & Short after certain days’, case study I](image)

![Figure IV.3 - Innovation: Daily ROI for S&P500 index and Profit for strategy ‘Top-10 Companies & Short after certain days’, case study I](image)

![Figure IV.4 - Popularity: Daily ROI for S&P500 index and Profit for strategy ‘Top-10 Companies & Short after certain days’, case study II](image)

![Figure IV.5 - Innovation: Daily ROI for S&P500 index and Profit for strategy ‘Top-10 Companies & Short after certain days’, case study II](image)
innovation by filtering tweets based on words with the highest fitness score. All strategies used in second case scenario surpass the ones from the first case scenario. For the Innovation study, the highest ROI obtained was 10.8% for the first case study and 11.8% for the second case study, both using strategy “Top-10 Companies & Short after certain days” and selling stocks after 2 days its purchase. For the Popularity study, the highest ROI obtained was 3.5% for the first case study and 5.5% for the second case study, both using strategy “Top-10 Companies & Short after certain days” and selling stocks after 2 days its purchase. This shows that popular and innovative words influence had a fast impact when characterizing a company’s innovation.

V. CONCLUSIONS

In this study it was tested the influence of words on Twitter when characterizing an innovative and popular company, listed in the S&P500 index. To perform this task it was created a dictionary of popular and innovative words, to filter tweets with relevant information about company’s popularity and innovation. In order to evaluate the information addressed in those tweets the system presented uses a Naïve Bayes algorithm, with a tested accuracy of 70.6%, to calculate the sentiment polarity of each tweet so it can be later correlated with the stock market movements and understand the impact of all words included in the innovative and popular dictionary. Two case studies were analysed in this work, differing in the use of a Genetic Algorithm (second case study) that, based on word’s features, helped to increase the profits obtained in the first case study, with a best case of 10.8% ROI, by outputting the words that better characterized companies in the second case study, with the best case of 11.8% ROI, during the simulation period considered. Both best cases of ROI were obtained using the long strategy “Top-10 Companies” and the short strategy “Short after certain days”, achieving the highest profits when selling company’s stocks after 2 days its purchase. From the results obtained it is possible to conclude that the popular and innovative words selected for this study had a significant impact when characterising company’s popularity and innovation.

There are some points that could be improved as future work:

- Creation of a dictionary for Innovation and Popularity with more words than the ones considered in this study, to increase the sample of words and the accuracy when investing based on those dictionaries;
- Take into consideration user’s influence and correlate it with the most influential words for each subject, increasing the number of relevant information of analysis and possibly obtaining better results;
- Increase the period of time dedicated to the simulation in order to obtain a wider view of words’ influence;

REFERENCES


