Investing in Stock IPOs with Sentiment Analysis from Twitter optimized by Genetic Algorithms

Bruno Miguel Filipe Guilherme
Instituto Superior Técnico,
Universidade Lisboa.
bruno.mfguilherme@gmail.com

***Abstract*— This research looks at the predictive capability of social networks, e.g. Twitter, focusing on the financial community and the stock market. The goal of this study is to predict the stock price of companies that have offered IPOs. This is performed through the analysis of Twitter messages and the differences between the sentiment analyses of 8 algorithms applied to study Twitter sentences regarding the financial market. The above referred algorithms differ in the way tweets are selected, the community is chosen and events are detected through Twitter. In order to optimize results a genetic algorithm solution is considered with the purpose of determining the weights of the applied indicators to each sentiment algorithm. The proposed approach, lead to returns of 59% during the first negotiation days, and of 72% a week after.**

***Keywords*— Twitter, IPO, Stock Market, Sentiment Analysis, Genetic Algorithm, Prediction**.

I. Introduction

For the last years the importance of social networks has been increasing globally. In 2006 the launch of Twitter contributed for a lot of this growth and, although there are other alternatives Twitter is very popular amongst users. This social network has the particularity of having a specific number of characters in each sent message and other certain characteristics that have made it one of the most used platforms nowadays.

In early 2010's a lot of investigative studies cropped up on the capacity of social networks, like the prediction and detection of events, of which Twitter was one of the most studied. People's opinion had already been studied as a way of predicting presidential election results [1], and box office success of movies [2]. There are also several articles studying Twitter in regards of the financial area, in order to predict the market share prices [3] and the best time to invest. With this study the goal is to predict the evolution and behaviour of several assets on the instant they are listed in the market for the first time (the moment of Initial Public Offer, IPO).

The companies scheduled for IPOs are announced one or two weeks ahead of being listed, however their IPO might not happen or get postponed to a different date. Due to it being an investment made on a company's new beginning, there are no historical data to analyse and makes it hard to predict the evolution of an asset on the first day of negotiation or the near future. This type of offer happens when the shares of a company are sold for the first time to the general public. This investment is done with the goal of expanding the business and raising more capital. The technical analysis is characterized by the group of indicators applied along the way based in market price variations of the financial assets. It can be taken as a social psychology applied as a poll that rules the behaviour of a market user at a certain moment in time, or a probability analysis based on market price being a reflection of identified crowd behaviour that follow repetitive patterns. The two techniques complement each other, and although the data used for each one are very different, several analysts can correlate fundamental and technical data in order to give an efficient evaluation of the market. For example, it is possible to apply a fundamental analysis as base to choosing potential profitable companies and then make a technical analysis using sentiment analysis to correlate Twitter messages.

In this study different methods of analysing tweets will be used, as well as different algorithms that are used to classify them. The study of event detection is referred and the prediction behaviour at the moment after public offering of different assets through tweets, and also the application of 8 different algorithms differentiated by concrete examples of "real" sentences (taken from Twitter) in the context of the financial market. The main contributions of this research paper are: 1- the creation of new methods for the collection of tweets; 2-predicting IPO prices from 7 to 30 days after the company is listed on the market; 3- an adaptive system with genetic algorithms capable of automatically identifying which company to invest in

This paper is organised as follows; in Section 2 the related work is discussed. Section 3 describes the proposed solution architecture with GA. In Section 4 the sentiment analysis algorithm tools were discussed. Section 5 describes the case studies and results. Section 6 draws the conclusions.

II. Related Work

With the increase popularity of social networks online communication has evolved quickly, users can now promptly create and share content easily, allowing for the interpretation of trends and discourses in society.

A. Twitter

In the landscape of social media Twitter is a popular platform, providing valuable information that is researched and analysed by various studies. The following image shows the process of studying a social network, particularly Twitter. More than 85% of topics mentioned in Twitter are news headlines or concerning environmental disasters [4]. Real-time updates allow for swift communication between users in different countries and locations, it has also become a fast and effective tool for the detection of disastrous events or terrorist incidents, communicating with the outside world and, in extreme cases, requesting assistance [5], [6].

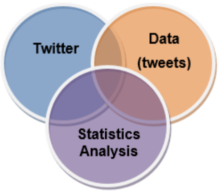


Fig. 1 - Power of Social Networks

In 2010, Huberman referred to the evolution of social networking and how it is possible, through the analysis of messages exchanged between users, to predict a multitude of topics including consumer's satisfaction regarding different products and election results forecast, showing that Twitter in particular is a very strong indicator of future results. Huberman aimed to predict a film's success before its premiere and his work spearheaded future studies associating Twitter to the financial market [2].

In 2012 Twitter created *cashtags*, using a dollar symbol ($) prefix users could better filter results about stocks and companies. However, some users still use the symbol of the shares without the *cashtag* or with the more common hashtag, which makes data collection difficult and less efficient for various surveys.

Stanford NER is a Java implementation of a Named Entity Recognizer, which labels sequences of words in a text that are the name of things (e.g., people and company names). It has been used in a number of studies with the objective of filtering through tweets and recognizing stocks and companies' names. Ritter in 2011 and Li in 2012, made use of Stanford NER comparing it to their own NER programs, which showed more effective results than the former, though only in a smaller scale [7], [8].

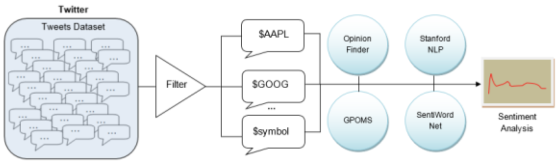


Fig. 2 - General Procedure for Sentiment Analysis

In 2011, Zhang et al [3] published the most important work at the time in predicting stock market indexes such as DJIA, NASDAQ, S&P 500 through Twitter, including a special mention to the previous mentioned study by Huberman. It showed that, on average, there was more than twice the number of positive tweets than negative ones, suggesting people send more messages when they are feeling more positive.

Another work often cited is the survey conducted by Bollen et al [1], wherein he analyses the accuracy of the stock market through moods stated on Twitter with the help of the OpinionFinder tool [9], also used in other studies Wilson et al [10] and Riloff and Wiebe [11], which classifies the moods as positive or negative; and the GPOMS algorithm that records the levels of six states of mind (happiness, kindness, alertness, sureness, vitality and calmness). The last phase of the study used a Self-Organizing Fuzzy Neural Network (SOFNN) algorithm [12] to test the hypothesis of DJIA prediction accuracy when the forecast model includes mood measures.

B. Initial Public Offerings (IPOs)

Initial Public Offerings (IPOs) are a type of public offering that occurs when a company first issues shares for sale to the general public in the stock exchange. This process is at the discretion of the company itself and tends to follow the same steps to come to its conclusion. The following diagram aims to demonstrate the steps that need to be taken for an IPO offer.

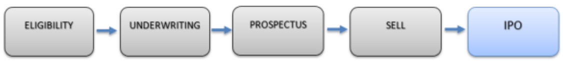


Fig. 3 - Process Steps of Initial Public Offering

In 2002, Ritter and Welch [13] considered that excessive enthusiasm among retail investors might explain the high returns on the first day and low long-term returns for an IPO. Their work is focused on the three main areas of research regarding IPOs: reasons for the IPO, pricing and allocation of shares, and long-time performance. The wide variation in the number of IPOs suggests the most important factor in the decision to go public is the market conditions at the time of offer. Long-time performance is the most controversial aspect in IPO research. While Ritter and Welch [13] favour a behavioural point of view as the determining factor, their results are sensitive/specific to both their methodology and the chosen time period.

In 2006, Corneli and Goldreich [14] studied whether post-IPO prices are influenced by small investors and were intent in determining if these investors should be considered "irrational investors". In this research project the price of the grey market is the indicator of small investor's, while older studies' objectives where to identify small investors' behaviour, specifically retail investors, in a more indirect way. Those analysis were focused on the relationship between small investors and bookbuilding investors [15].

Cornelli, Goldreich and Ljungqvist aimed to identify the investor sentiment post-IPO. They collected data from 486 companies that went public in 12 different European countries between November 1995 and December 2002. They estimated that excess optimism by grey market investors are the main cause of these IPOs being traded on the first day at a price up to 40.5% higher than it would have been in the absence of optimistic feelings. Over the subsequent 12 months of trading, as unrealistic expectations give way to reality, prices fall. They concluded that an excess of optimism by small investors make them have unrealistic expectations about the company and are therefore willing to pay more than the basic price for an IPO.

In an attempt to forecast IPOs in 2008, Zhang, Liu and Sherman [16] analysed the importance of the media in the IPO value. It was concluded that the media has a direct relation to the IPO price; it was observed that the good opinion of the media in regards to a specific company usually raises the value IPO or it suffers a minor adjustment, in order to obtain higher profits and high returns on the first day. A positive view of the company increases the demand for the stock leading to a rise in prices.

In 2013, Loughran and McDonald [17] studied an IPO's first day returns based on the offer price revisions, volatility and the S-1 document of the company, and concluded that a company whose S-1 document is a text with a "high degree of uncertainty, has the largest absolute first day and offer price revisions and subsequent volatility returns." They believe that higher uncertainties regarding an IPO leads to higher yields in the first day and greater volatility of resale, taking into account the sample IPOs from 1997 and 2010. This study also examines the sentiment in the S-1 documents and show that IPOs that use a higher rate of negative words tend to have higher absolute revisions in offer prices, more consistent with the hypothesis of product information.

C. Twitter and IPOs' forecast

In 2015, Jim Liew and Garret Wang [18] published the latest version of the study in which they analyse the cross between investor sentiment and IPO behaviour on the first day of trade. This is the first work correlating between Twitter and IPOs, and it concludes that there is a correlation between investor sentiment in the days before the IPO and the IPO return on the first day of trading.

The study tested three hypotheses. Hypothesis 1 considered the sentiment of the average retail investor for each IPO from the opening to the closing on the first day of trading in relation to the first day returns from opening price to closing price. Hypothesis 2, the sentiment of the average investor for each IPO in the 1st, 2nd and 3rd day before the first day of trading in relation to the first day returns from opening price to closing price. Hypothesis 3 was designed to correlate the sentiment of the average investor for each IPO from opening to 20, 30 and 60 minutes following the first day of training with the returns of the first day at 21, 41 and 61 minutes after the initial offer until the closing. The first hypothesis showed more consistent results and a positive and significant correlation, unlike hypotheses 2 and 3.

Tamy Kwan [19] uses the volume of tweets as an indicator to forecast IPOs' returns. Kwan analyses the evolution of certain IPOs depending on the amount of references on the company. Due to the difficulty of collecting older tweets, the data collected for analysis was obtained from older databases already used in various studies. It concluded that there is a greater number of tweets on the first trading day, which may mean that users like to talk about IPOs with better performance, and during the negotiation process, which could affect the closing price. In regards to the days leading up to the trading day, it showed that there is no predictive relationship between the number of tweets and IPO performances. Note that in this study the sentiment analysis algorithm evaluated tweets as positive vs negative, contributing to unreliable results. Table 1 summarizes some of the most relevant existing solutions in the area of IPOs and Twitter specified according to several parameters.

Reference	Study Focus	Year	Main Methods	Test Period	Market (Finance)	Algorithm Performance
J. Bollen [1]	Twitter and stock market	2010	OpinionFinder and GPOMS	03/2008 - 12/2008	DJIA	SOFNN: 86.7%
M. Hubberman [2]	Twitter and prediction	2010	LingPipe and DynamicLMClassifier	09/2009 - 01/2010	Cinema ticket.	N/A
X. Zhang [3]	Twitter and stock market	2011	Number of tweets and followers	03/2009 - 09/2009	DJIA, NASDAQ and S&P 500	N/A
J. Ritter [13]	IPO	2002	Linear Regression	1980 - 2001	US Market	N/A
F. Cornelli [14]	IPO	2006	Math and static methods	11/1995 - 12/2002	Grey Market	N/A
L. X. Liu [16]	IPO	2008	Linear Regression	01/1980 - 12/2004	US Market	N/A
T. Loughran [17]	IPO	2013	Math and static methods	1997 - 2010	US Market	N/A
J. Liew [18]	Twitter and IPO	2015	OpinionFinder, GPOMS and iSENTIUM LLC	01/2013 - 12/2014	NYSE, NASDAQ	Linear Regression: N/A
T. Kwan [19]	Twitter and IPO	2015	Twitter API and Naïve Bayes Classifier	06/2009 - 12/2009 01/2014 - 05/2014	N/A	N/A

Table 1 – State of art overview.

III. Solution Architecture

This chapter describes the proposed solution architecture. The application of various sentiment analysis algorithms is explained in more detail in the next chapter.

Gnip [20], Twitter's enterprise API platform, makes available to the general public a license for access to the database containing all tweets since 2006, and also full access to real-time tweets occurring all around the globe at any time. However, this license is quite expensive and even for the purpose of advancing a dissertation it wasn't possible to obtain one.

In order to overcome this major barrier a methodology was designed based on the idea of "six degrees of separation", wherein it would be possible to access the last two years of tweets of most users by following the connections between users through six levels. As for the technology used to develop this system, an interface was implemented running the REST API 1.0 [21] through Twitter4J [22].

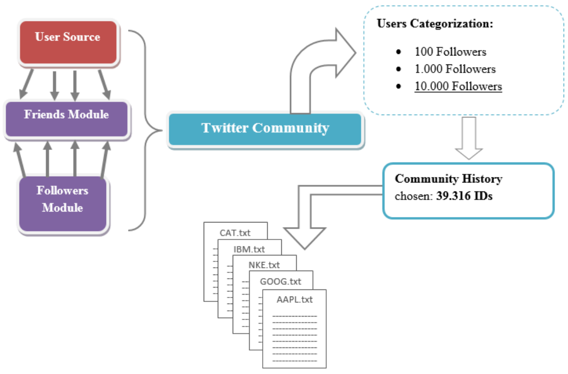


Fig. 4 - Representation of the implemented interface

The system is composed of six basic components whose purpose is described below:

- Source User: main user who serves as a starting point to the Twitter community it generates through the friends the user follows.
- Friends Component: set of users (each identified by a unique ID) that are followed by the source user, called friends.
- Followers Component: set of users who follow a specific user ID, called followers. In the diagram it shows followers of friends.
- Categorization of users: in order to restrict to a set of users, they are categorized according to the number of followers they represent. In theory, the higher the amount of followers, the higher the level of importance.
- History of users: because of Twitter limitations, restricted to the last 3200 tweets of a particular user.
- Active filtering: the total number of saved tweets of the various users are filtered and grouped by asset symbol quoted on the stock market. It is possible to filter by symbol or name of the company.

In broad terms, the flow of data inside the API starts with the system requesting the user parameters, then the optimization algorithm is executed, and lastly a recommended investment is created (see fig. 5). More specifically, the whole process runs as follows:

1. The user starts by setting the parameters (language, location, keywords, users) of the search concerning the specific asset, for example a search regarding Apple, Inc. would contain the parameters: keywords:$APPL, #APPL, iPhone, Apple Inc; language: English; location: USA;
2. Then the system applies a set of algorithms to classify the sentiment of the tweets;
3. After this process the genetic algorithm starts by defining a random set of individuals, each corresponding to different weight classes within the various algorithms in order to determine which algorithm brings about optimal solutions;

4. In order to evolve, the system scores each classification by the number of hits and corresponding fitness value;
5. When the genetic algorithm converges to a final solution the system executes the investment simulator again using present time data in order to offer a safe investment for the day;
6. Every week the system is executed in order to update the latest data as well as upcoming IPOs for future evaluation.

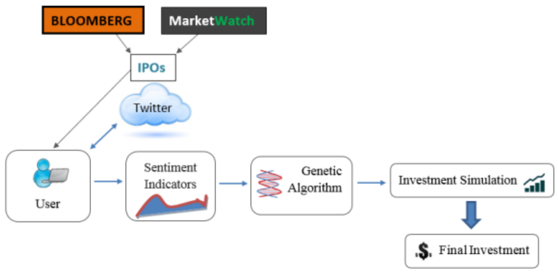


Fig. 5 - Project Flowchart

A. *Data Processing Component IPOs*

The component responsible for the processing the IPOs data uses information from the Bloomberg Professional application resource [23] and information obtained from the website MarketWatch [24], dedicated to IPO events, the website can identify all scheduled IPOs.

For this study companies with scheduled IPOs since October 2015 were analysed. The Research Component has the responsibility of accessing recent tweets by word, and an optimization system was implemented to search by symbol and company name efficiently and without duplicate data. Then, of the IPOs that have been confirmed, the tweets are analysed and recorded in the database by the Registration Component. The message is saved along with its evaluation according to the different sentiment algorithms in use. In case there is a need to implement new releases or new analysis algorithms, there is a text field in the data base that enables an easier comparison between the results and the algorithm's development process.

After several months it is possible to compile and analyse a historical database of tweets relating to various assets.

B. *Optimization Module*

This component is responsible for optimizing the sentiment indicators used and the configuration of the system so it is able to classify and provide certain market models in order to make predictions for the real world. To optimize the system, genetic algorithms were chosen based on their versatility and prior experience.

Starting with the representation of the chromosome, an individual in the population is represented by a real-value vector structure in which each element corresponds to the weight, the importance given to each specific sentiment algorithm in the classification equation. In addition to its weight, the chromosome also can adopt a gene called "observation days" so it can consider a long or a short sample data. The "observation days" variable represents the most important period for the IPO forecast, which is the days preceding the date of the IPO and therefore tweets with more impact on the performance of stock market value. Each analysis algorithm has a specific weight within the classification model. The classifier is given by the following equation:

$$S = \sum_{i=0}^{N} W_i \,.\, Score(X, i) \qquad (1)$$

$$0 \le W_i \le 1 \qquad (2)$$

$$\sum_{i=0}^{N} W_i = 1 \qquad (3)$$

Where:

- W_i is the weight associated with the *i* analysis algorithm;
- Score(X, *i*) is the resulting value obtained by *i* analysis algorithm for asset X.

After the algorithm's optimization process, which results in the classification equation, when the set of weights are properly balanced, all active data that is quoted in the market for more than 30 days is classified. In Table 2 is presented the chromosome representation, the sentiment algorithms presented in the genes are explained in detail in section 4.

Observation days	AFFIN	TextBlob	SWN	NLTK	Senti Strength	Unigrams & Bigrams
[1, 30]	[0, 1]	[0, 1]	[0, 1]	[0, 1]	[0, 1]	[0, 1]

Table 2 – Chromosome representation

C. *Investment Simulation Module*

In an investment simulation system it is necessary to create a dataset of observations in accordance with the equation to classify a person. This management module is used by the genetic algorithm to classify each chromosome and create a training simulation to the current situation.

In order to implement a realistic experience it is necessary to divide the dataset in a subgroup of training and a subgroup of test. In order to assess the ability of each individual, the stock behaviour of each asset was recorded within 1, 7 and 30 days post-IPO, designated by dataset module. The objective at this stage is to reach the maximum number of hits to the selected dataset, making it possible to arrive at the best solution proposed by the optimization algorithm.

Next, it is shown the structural example implemented on the dataset.

Dataset Assets	Day 1	Week 1	Month 1	Month 2
#1	1	0	0	0
#2	0	0	1	0
…	…	…	…	…
#N	1	1	1	0

Table 3 – Example of the dataset used

It is designated by 1 or 0 the stock of a company that after X days had an increase or decrease in price, respectively, for the N assets selected in the training period.

As shown in equation number (2) it is subsequently implemented a mapping function to the discrete space of the resulting value S of the equation as shown below:

$$S \in \mathbb{R} \rightarrow Z = f(S) \rightarrow Z \in \mathbb{N} \qquad (4)$$

After performing the mapping function to the value S, the number of hits is computed by comparing the result with the previously registered data set. The optimal solution is the solution that maximizes the total number of hits (Z function).

IV. Sentiment Algorithms

The main goal of this research is the application of text mining algorithms to understand the sentiment of twitter texts. In order to achieve this goal eight sentiment analysis algorithm tools were implemented. This type of analysis is now widely studied and there is a wide range of possible solutions. Initially the implemented algorithms were based on rating punctuated words:

- *TextBlob*
- *Unigrams and Bigrams classification*
- *AFFIN 1.0*
- *SentiStrength*

A major limitation of this kind of algorithm was that they were programmed to give some sort of opinion regardless of context, because the words are analyzed in isolation and have a fixed rating. These algorithms don't have the ability to perceive subjectivity beyond polarity and so, to overcome this limitation and evolve the analysis of the simulated system, the following algorithms were implemented:

- *Stanford CoreNLP*
- *POS tagging – part of speech tagging*
- *SentiWordNet 3.0*
- *NLTK – Natural Language Toolkit*

TextBlob [25] is a Python library for data processing. With the use of an API, a set of features called Natural Language Processing is provided, such as the categorization of words in the context of the sentence, the sentence extracting the name, sentiment analysis, speech detection, among others.

The analysis algorithm Uni and Bigrams, implemented in Java, aims to classify a sentence through unigrams and bigrams, without the ability to evaluate emoticons, and the resulting value given by:

$$\sum Unigrams(sentence) + Bigrams(sentence) \tag{5}$$

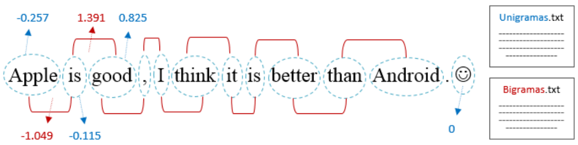


Fig. 6 - Operating mode for Uni and Bigrams classification

Very similar to the previous algorithm in the way it works, the sentiment analyser AFFIN [26], [27], implemented in Python classifies each word entered, with the resulting value being the sum of the value given to each word. It has the particularity of identifying emoticons.

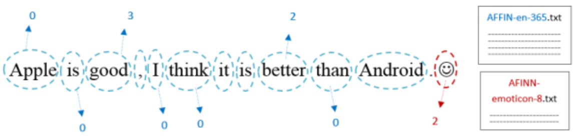


Fig. 7 - Operating mode for AFFIN algorithm

SentiStrength [28], [29] was applied to the comments of the MySpace social network and accurately presented sentiment rating values able to predict positive emotions in the order of 60% and negative emotions with an accuracy of 72%. It was designed in Python and aims to extract the strength of feeling of informal text, used in public messages between users on social networks, using new methods to explore the diversity of grammar and spelling styles found on social networks.

Stanford Core NLP offers a set of tools for the analysis of natural language. It can recognize the root form of words, grammatical classes, recognize names of people, companies, normalizing dates, indicate the noun in a sentence and the feeling it translates. This tool constructs a representation of a sentence based on its structure. It calculates the feeling based on how the words make up the meaning in long sentences.

An analysis and classification program in Java, SentiWordNet version 3.0 is a lexical resource explicitly designed to support the application of sentiment analysis and classification of opinions. The revolutionary aspect of this algorithm is that it analyses a word taking into consideration the word's class including noun (NN), adjective (JJ), adverb (RB) and verb (VB). To this end, the initial analysis is complemented by the POSTaggerTool:

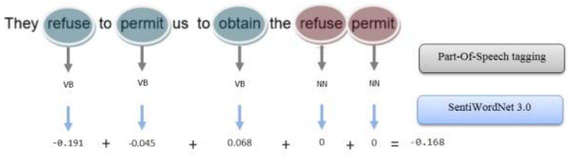


Fig. 8 - Tag Words Method

NLTK is a Python tool widely used in word processing and presents itself as one of the most complete in this area. This algorithm requires training to use. It showed excellent results even when tested using difficult sentences or tricks that would lead to wrong results in many analysis tools.

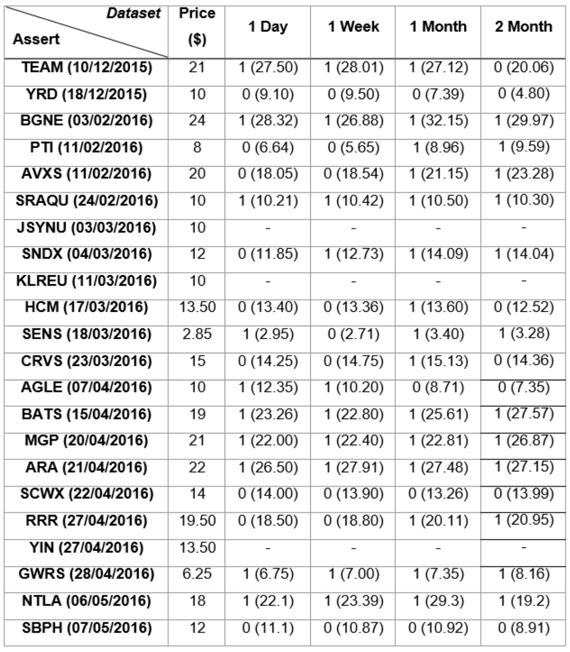
The incorrect semantics found on Twitter messages influence the way Stanford CoreNLP operates, leading it to be the worst performing algorithm, which lead it to being dropped from the final assessment, as was SentiWordNet. Social network messages often have grammatical errors and abbreviations which is a problem for any text classifier, and the categorization of the grammatical class of each sentence is strongly conditioned by it.

V. EXPERIMENTAL RESULTS

In this section, two case studies related to IPO investment are presented. In case 1 the total number of tweets that mention specific assets are taken into account, thus identifying its popularity. In case 2, sentiment analysis genetic algorithm determines weights to predict the behaviours of each post-IPO asset.

The effect of stock market volatility in 2016 reduced the number of IPOs launches, consequently there was a limited sample in comparison with the set of observed data. According to Bloomberg MarketWatch analysis even though there were 208 companies with possible IPOs scheduled, only 26 of which were realized.

Table 4 presents the set of assets identified by symbol and date of release, followed by the initial stock price at different times (1, 7, and 30 days). The company Jensyn Acquisition Corp (JSYNU), KLR Energy ACquisiton Corp (KLREU) and Yintech Investment Holdings Ltd (YIN) had very little significant volumes of shares and their prices changes after stock market listings is practically null (~0.50 quotation price), therefore their values were not included.



Dataset **Assert**	**Price ($)**	**1 Day**	**1 Week**	**1 Month**	**2 Month**
TEAM (10/12/2015)	21	1 (27.50)	1 (28.01)	1 (27.12)	0 (20.06)
YRD (18/12/2015)	10	0 (9.10)	0 (9.50)	0 (7.39)	0 (4.80)
BGNE (03/02/2016)	24	1 (28.32)	1 (26.88)	1 (32.15)	1 (29.97)
PTI (11/02/2016)	8	0 (6.64)	0 (5.65)	1 (8.96)	1 (9.59)
AVXS (11/02/2016)	20	0 (18.05)	0 (18.54)	1 (21.15)	1 (23.28)
SRAQU (24/02/2016)	10	1 (10.21)	1 (10.42)	1 (10.50)	1 (10.30)
JSYNU (03/03/2016)	10	-	-	-	-
SNDX (04/03/2016)	12	0 (11.85)	1 (12.73)	1 (14.09)	1 (14.04)
KLREU (11/03/2016)	10	-	-	-	-
HCM (17/03/2016)	13.50	0 (13.40)	0 (13.36)	1 (13.60)	0 (12.52)
SENS (18/03/2016)	2.85	1 (2.95)	0 (2.71)	1 (3.40)	1 (3.28)
CRVS (23/03/2016)	15	0 (14.25)	0 (14.75)	1 (15.13)	0 (14.36)
AGLE (07/04/2016)	10	1 (12.35)	1 (10.20)	0 (8.71)	0 (7.35)
BATS (15/04/2016)	19	1 (23.26)	1 (22.80)	1 (25.61)	1 (27.57)
MGP (20/04/2016)	21	1 (22.00)	1 (22.40)	1 (22.81)	1 (26.87)
ARA (21/04/2016)	22	1 (26.50)	1 (27.91)	1 (27.48)	1 (27.15)
SCWX (22/04/2016)	14	0 (14.00)	0 (13.90)	0 (13.26)	0 (13.99)
RRR (27/04/2016)	19.50	0 (18.50)	0 (18.80)	1 (20.11)	1 (20.95)
YIN (27/04/2016)	13.50	-	-	-	-
GWRS (28/04/2016)	6.25	1 (6.75)	1 (7.00)	1 (7.35)	1 (8.16)
NTLA (06/05/2016)	18	1 (22.1)	1 (23.39)	1 (29.3)	1 (19.2)
SBPH (07/05/2016)	12	0 (11.1)	0 (10.87)	0 (10.92)	0 (8.91)

Table **4** – Dataset used for training and testing

A. *Case study nº1: Assets Popularity and Event Detection*

In this subchapter, case study 1, the investment returns will be analyzed taking into consideration the popularity of each asset according to the number of tweets that mention a specific asset.

An important point noted in this IPO analysis is that the messages observed served mostly as advertising of IPOs, referencing which would occur in the coming weeks, so, in the case of simple announcements or news, they do not express feeling and are merely informative, demonstrating the capacity of Twitter in regards to event detection. The following graph show the volume of tweets in relation with a set of observed IPOs.

Figures 9 and 10 show the evolution of tweets over time for a set of six stocks. The date the stock started trading in the stock exchange is very clearly marked by peaks in the chart and with the ☆ symbol. As mentioned previously, assets that did not achieve a significant number of tweets during the pre-negotiation period suffered virtually no changes since the initial stock price on the stock exchange (marked by 🛇, KLREU and JSYNU had only a few messages, insufficient to identify or advertise the event itself). Assets that observe a sharp positive slope (first day of trade) will suffer some kind of change in its price during the next few days. In figure 10, the assets with larger number of tweets during that time are shown.

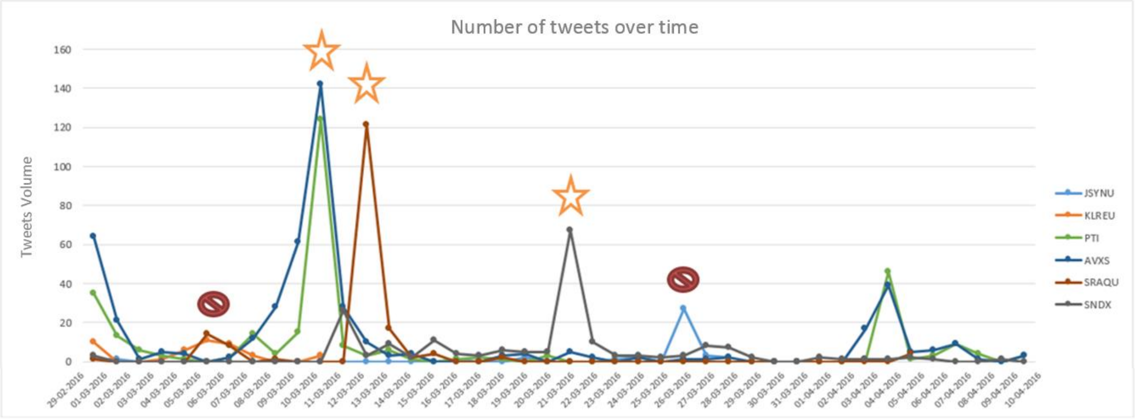


Fig.9 - Number of tweets over time: event detection

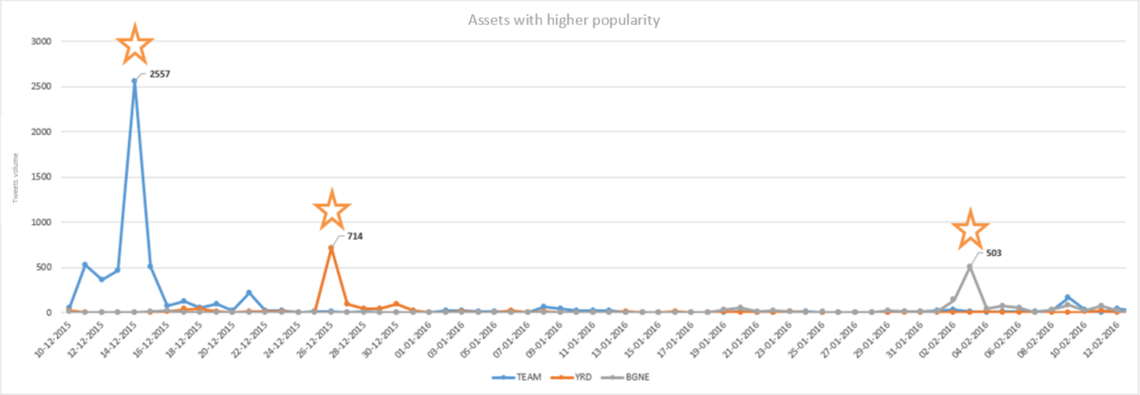


Fig.10 - Assets with higher popularity: Atlassian Inc., Yirenday Ltd, Beigene Ltd.

It is possible to observe a large gain in figure 10 concerning companies that were popular and had a large number of tweets; however the growth rates are unpredictable, unlike assets that aren’t much talked about and show a gradual increase of tweets.

Taking into consideration the results of the case study, the best strategy is to buy the asset whose evolution of number of tweets is gradual and that are only somewhat popular. In assets that had no change in tweets, there were no variation in price and they would not make a good investment opportunity. Chinese companies are much too volatile and unpredictable in their prices variations to be a good investment. In cases where the assets are very popular and there is an overabundance of tweets, there is sometimes an asset overvaluation, and, according to the data analysed in this case study, their behaviour is unpredictable.

B. *Case study nº2: Genetic Algorithm Optimized Solution*

In this case study, the multiple tweets about IPOs and companies that went public was studied using different sentiment-based algorithms.

It analysed the optimal solution depending on different time periods of the dataset used.

In order to understand the practical results of the case study two rates were set based on observations made:

- **Hit rate**: given in percentage points, it refers to the ability of the simulation system to determine the amount of investment to be made in accordance to the asset behaviour. It is given by:

$$Hit\ Rate\ (\%) = \frac{Total\ number\ of\ hits}{Total\ number\ of\ assets} \times 100 \qquad (6)$$

- **Gain rate:** percentage of the value gained in comparison to the value invested. It is given by:

$$Gain\ Fee\ (\%) = \frac{Price_{final} - Price_{initial}}{Price_{initial}} \times 100 \qquad (6)$$

The training process and optimization module tests on the best solution seeks included a set of data 22 assets and each one is formed in a random manner so as to vary the maximum each set so as to overcome the problem of low IPOs confirmed sample.

The following graph shows the gain variations for each asset taking into account the actual behaviour of the shares at different times.

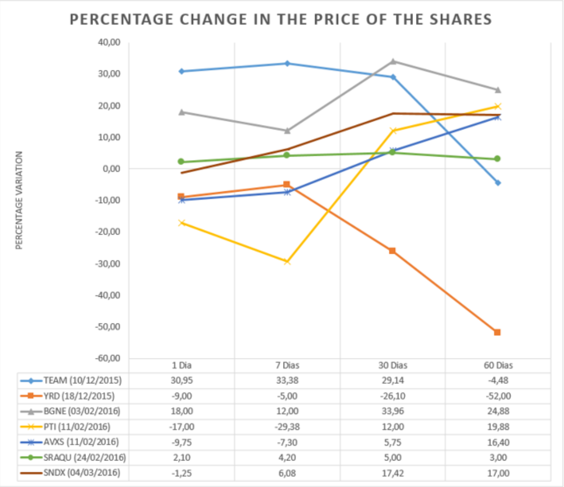


Fig.11 - Graphic representation of the percentage change in earnings per active

Real situation				
	Gain (%)			
	1 Day	7 Days	30 Days	60 Days
TEAM (10/12/2015)	30,95	33,38	29,14	-4,48
YRD (18/12/2015)	-9,00	-5,00	-26,10	-52,00
BGNE (03/02/2016)	18,00	12,00	33,96	24,88
PTI (11/02/2016)	-17,00	-29,38	12,00	19,88
AVXS (11/02/2016)	-9,75	-7,30	5,75	16,40
SRAQU (24/02/2016)	2,10	4,20	5,00	3,00
JSYNU (03/03/2016)	0,00	0,00	0,00	0,00
SNDX (04/03/2016)	-1,25	6,08	17,42	17,00
KLREU (11/02/2016)	0,00	0,00	0,00	0,00
HCM (17/03/2016)	-0,74	-1,04	0,74	-7,26
SENS (18/03/2016)	3,51	-4,91	19,30	15,09
CRVS (23/03/2016)	-5,00	-1,67	0,87	-4,27
AGLE (07/04/2016)	23,50	2,00	-12,90	-26,50
BATS (15/04/2016)	22,42	20,00	34,79	45,11
ARA (21/04/2016)	20,45	26,86	24,91	23,41
MGP (20/04/2016)	4,76	6,67	8,62	27,95
SCWX (22/04/2016)	0,00	-0,71	-5,29	-0,07
RRR (27/04/2016)	-5,13	-3,59	3,13	7,44
YIN (27/04/2016)	0,00	0,00	0,00	0,00
GWRS (28/04/2016)	8,00	12,00	17,60	30,56
NTLA (06/05/2016)	22,78	29,94	62,78	6,67
SBPH (07/05/2016)	-7,50	-9,42	-9,00	-25,75

Table **5** – Gain variation at different times

Table 6 shows the resulting value of investment due to the positive or negative change based in the initial amount investment.

LONG	GAIN (€)			
SHORT	1 Day	7 Days	30 Days	60 Days
TEAM (10/12/2015)	309 523,81 €	333 809,52 €	291 428,57 €	- 44 761,90 €
YRD (18/12/2015)	- 90 000,00 €	- 50 000,00 €	- 261 000,00 €	- 520 000,00 €
BGNE (03/02/2016)	180 000,00 €	120 000,00 €	339 583,33 €	248 750,00 €
PTI (11/02/2016)	- 170 000,00 €	- 293 750,00 €	120 000,00 €	198 750,00 €
AVXS (11/02/2016)	- 97 500,00 €	- 73 000,00 €	57 500,00 €	164 000,00 €
SRAQU (24/02/2016)	21 000,00 €	42 000,00 €	50 000,00 €	30 000,00 €
JSYNU (03/03/2016)	- €	- €	- €	- €
SNDX (04/03/2016)	- 12 500,00 €	60 833,33 €	174 166,67 €	170 000,00 €
KLREU (11/02/2016)	- €	- €	- €	- €
HCM (17/03/2016)	- 7 407,41 €	- 10 370,37 €	7 407,41 €	- 72 592,59 €
SENS (18/03/2016)	35 087,72 €	- 49 122,81 €	192 982,46 €	150 877,19 €
CRVS (23/03/2016)	- 50 000,00 €	- 16 666,67 €	8 666,67 €	- 42 666,67 €
AGLE (08/04/2016)	235 000,00 €	20 000,00 €	- 129 000,00 €	- 265 000,00 €
BATS (15/04/2016)	224 210,53 €	200 000,00 €	347 894,74 €	451 052,63 €
ARA (21/04/2016)	204 545,45 €	268 636,36 €	249 090,91 €	234 090,91 €
MGP (20/04/2016)	47 619,05 €	66 666,67 €	86 190,48 €	279 523,81 €
SCWX (22/04/2016)	- €	- 7 142,86 €	- 52 857,14 €	- 714,29 €
RRR (27/04/2016)	- 51 282,05 €	- 35 897,44 €	31 282,05 €	74 358,97 €
YIN (27/04/2016)	- €	- €	- €	- €
GWRS (28/04/2016)	80 000,00 €	120 000,00 €	176 000,00 €	305 600,00 €
NTLA (06/05/2016)	227 777,78 €	299 444,44 €	627 777,78 €	66 666,67 €
SBPH (07/05/2016)	- 75 000,00 €	- 94 166,67 €	- 90 000,00 €	- 257 500,00 €

Table **6** – Gain in € per active with the amount of € 1 million invested

Taking into account equations (8) and (9), the results of the simulation system and of a Long & Short strategy were analysed. With that in mind, from an investment value of 1 million per asset, table 8 shows a comparison between the investment simulator and the investment strategies (always choosing the best Long & Short strategy and get the highest possible profit). If the system was ideal and obtained a success rate of 100%, the investment in each earning figures would be the ones shown in the Total (Possible) line, followed by different Long & Short strategies.

Unlike the ideal returns for each asset, the calculations necessary in order to assess the best Long & Short strategies are shown.

	1 Day	7 Days	30 Days	60 Days
TOTAL (Possible):	1 815 676,02 €	1 611 998,59 €	2 367 768,37 €	2 872 779,99 €
SHORT	553 689,46 €	630 116,80 €	532 857,14 €	1 203 235,45 €
LONG	1 564 764,34 €	1 531 390,33 €	2 759 971,05 €	2 373 670,18 €
Investment Value:	1 000 000,00 €			
Gain (%) - Total	181,57	161,20	236,78	287,28
Strategy SHORT	55,37	63,01	53,29	120,32
Strategy LONG	156,48	153,14	276,00	237,37

Table **7** – Ideal Gains of different strategies

As can be seen in Table 7, the investment gain on the first trading day is higher when compared to other times (a 181.57% gain). A Short strategy has a small gain (55.37%), which contradicts the common investment option for post-IPO shares, logically contradicted by Long strategies that always had a higher gain.

These strategies were heavily influenced by cases of extreme success and increase in value such as Atlassian Corporation PLC (NASDAQ:TEAM).

In Table 8 we present the solution resulting from the genetic algorithm during the training process and show the best results.

Prediction Period	No. of days	AFFIN	TextBlob	SWN	NLTK	Senti Strength	Uni & Bigrams
1 Day	1	1.8 %	8 %	1.6 %	1.7 %	84.7 %	2.2 %
7 Days	3	0.1 %	4.7 %	0.9 %	0.7 %	92.8 %	0.8 %
30 Days	27	72 %	5.9 %	4 %	9 %	4.1 %	5 %

Table **8** – Optimized solution

Taking into consideration the solution presented, the weights of sentiment analyzers are, for the most part, distributed to the algorithms that are incapable of interpreting semantics. With 72% for AFIIN in a 30-day prediction, and in the order of 90% the SentiStrenght algorithm in a short term prediction.

On table 9 the profits of the application developed during the test phase during the different temporal goals defined are shown.

The system shows a success rate of approximately 50% in the test phase for the first day, and of 68% for the week. On the long run, the success rate was higher than expected, which can be explained by the selection of training and test models. These were selected randomly out of the set of IPOs observed through the research; however, this is not an optimal solution. All records refer to the same period of time and, therefore, may correlate between themselves.

	1 Day	7 Days	30 Days	60 Days
Investment simulator	589 734,34 €	725 809,52 €	690 823,46 €	1 265 749,02 €
Gain (%)	58,97	72,58	69,08	126,57
Hit rate(%)	50%	68%	90%	90%

Table **9** – System Simulation Gain

The investment simulator can track the IPO behaviour forecast more effectively in the early days of trading, and assures an increase in investment of 65% with up to 70% accuracy on occasion.

If it were possible to obtain a more representative sample of IPOs, the training and testing process would be improved, however due to situation of the market in early 2016, the scheduled IPOs didn't come to fruition and it took time for new assets to enter the market. Beyond that, if it was at all possible to access data at different times, the market sample would be much more significant leading to greater certainty and accuracy in the simulation system.

The fact that an IPO is an event with unique characteristics leads to significant uncertainty in initial negotiations and very different variations. The focus of the data analysis was Twitter, which showed potential to be a possible platform for stock price forecasts.

VI. Conclusion and Future Works

This paper proposes an investment simulation system for assets listed for the first time in the market through the application of technical indicators in order to achieve the best possible gains/returns. To validate the proposed application/solution, the strategy was compared to the return results achieved with other strategies like the Long & Short, starting on November 2015 until May 2016. The preliminary results were not ideal, which lead to an in-depth study on the different analysis algorithms. The algorithm can be easily broadened and parameterized into another focus of market analysis e updated with new analysis techniques.

The simulation system has shown positive results and, through the applied indicators, it was possible to obtain interesting returns with a good success percentage, in the order of 70% for the first days after IPO negotiations. There is still a lot of study to be done when using this kind of Twitter prediction approach, because, beyond the sentiment extracted from sent messages, there are many other indicators present in each user and message that can improve the analysis. Besides that, the sentiment analysis tweet by tweet in a an isolated manner has shown inconclusive results, demonstrating that even for humans it's hard to classify several tweets as positive or negative because of lack of context and insufficient information contained in a message. However, the project developed has managed to create a comparison of several sentiment analysis algorithms focused on the market and taken from Twitter. Many algorithms were set up thinking of well-constructed sentences and produce good results with difficult and complex sentences, but in this study have been shown to be less effective because of sentence construction.

Due to the high data abstraction in the developed code, it is possible to expand on the solution reached. Below there are a few proposals on improvements on the present solution:

- Expanding the chromosome with more technical indicators. This expansion can be easily improved, all that is necessary is the desired indicator and to define the respective validation rules.
- Consideration of a parallel process in obtaining and classifying data in order to speed up the process.
- Programming webservices capable of using Twitter to obtain older data in order to overcome the historical difficulty/problem.
- Try and understand, through the applied Twitter interface, which are the most relevant indicators in the social network that contribute to the classification of a good tweet or a good user to analyze.
- Constructing and predicting the IPO index for the year.

References

[1] J. Bollen, Huina Mao and Xiaojun Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science,* vol. 2.1, pp. 1-8, 2011.

[2] B. A. Huberman and S. Asur, "Predicting the Future With Social Media," *Web Intelligence and Intelligent Agent Technology (WI-IAT) 2010 IEEE/WIC/ACM International Conference,* vol. 1, pp. 492-499, 2010.

[3] X. Zhang, H. Fuehres and P. A. Gloor, "Predicting Stock Market Indicators Through Twitter "I hope it is not as bad as I fear"," *Procedia-Social and Behavioral Sciences,* vol. 26, pp. 55-62, 2011.

[4] H. Kwak, C. Lee, H. Park and S. Moon, "What is Twitter, a Social Network or a News Media?," *Proceedings of the 19th international conference on World wide web. ACM,* pp. 591-600, 2010.

[5] T. Sakaki, M. Okazaki and Y. Matsuo, "Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors," *Proceedings of the 19th international conference on World wide web. ACM,* pp. 851-860, 2010.

[6] S. Vieweg, A. L. Hughes, K. Starbird and L. Palen, "Microblogging during two natural hazards events: what twitter may contribute to situational awareness," *Proceedings of the SIGCHI conference on human factors in computing systems. ACM,* pp. 1079-1088, 2010.

[7] A. Ritter, S. Clark, M. Etzioni and O. Etzioni, "Named Entity Recognition in Tweets: An Experimental Study," *Proceedings of the Conference on Empirical Methods in Natural Language Processing,* no. Association for Computational Linguistics, pp. 1524-1534, 2011.

[8] C. Li, J. Weng, Q. He, Y. Yao and A. Datta, "TwiNER: Named Entity Recognition in Targeted Twitter Stream," *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. ACM,* pp. 721-730, 2012.

[9] [Online]. http://www.cs.pitt.edu/mpqa/opinionfinderrelease/.

[10] T. Wilson, J. Wiebe and P. Hoffmann, "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis," *Proceedings of the conference on human language technology and empirical methods in natural language processing,* pp. 347-354, Association for Computational Linguistics, 2005.

[11] E. Riloff and J. Wiebe, "Learning Extraction Patterns for Subjective Expressions," *Proceedings of the 2003 conference on Empirical methods in natural language processing,* pp. 105-112, Association for Computational Linguistics, 2003.

[12] G. Leng, G. Prasad and T. M. McGinnity, "An on-line algorithm for creating self-organizing fuzzy neural networks," *Neural networks,* vol. 17, no. 10, pp. 1477-1493, Elsevier, 2004.

[13] J. Ritter and I. Welch, "A review of IPO activity, pricing, and allocation," *The Journal of Finance,* vol. 57, no. 4, pp. 1795-1828, Wiley Online Library, 2002.

[14] F. Cornelli, D. Goldreich and A. Ljungqvist, "Investor Sentiment and Pre-IPO Markets," *The Journal of Finance,* vol. 61, no. 3, pp. 1187-1216, Wiley Online Library, 2006.

[15] F. Cornelli and D. Goldreich, "Bookbuilding and Strategic Allocation," *The Journal of Finance,* vol. 56, no. 6, pp. 2337-2369, Wiley Online Library, 2001.

[16] L. X. Liu, A. E. Sherman and Y. Zhang, "The Role of the Media in Initial Public Offerings," *DePaul University and Hong Kong University of Science & Technology Working Paper,* 2009.

[17] T. Loughran and B. McDonald, "IPO First-Day Returns, Offer Price Revisions, Volatility, and Form S-1 Language," *Journal of Financial Economics,* vol. 109, no. 2, pp. 307-326, 2013.

[18] J. K.-S. Liew and G. Z. Wang, "Twitter Sentiment and IPO Performance: A Cross-Sectional Examination," *Available at SSRN 2567295,* 2015.

[19] T. Kwan, "Twitter Volume and First Day IPO Performance," 2015. [Online]. http://www.stern.nyu.edu/sites/default/files/assets/documents/2015%20Kwan%20paper.pdf

[20] "GNIP product," GNIP Inc., 2015. [Online]. https://gnip.com/.

[21] I. Twitter, "Documention, REST API," Twitter, Inc., [Online].

https://dev.twitter.com/overview/documentation. [Accessed 2016].

[22] "Twitter4J," [Online]. http://twitter4j.org/en/index.html.

[23] B. F. L.P, "Bloomber Professional," Bloomberg Finance L.P, [Online]. http://www.bloomberg.com/professional/remote-access-mobile/.

[24] "MarketWatch: IPO Calendar," MarketWatch, 2016. [Online]. http://www.marketwatch.com/tools/ipo-calendar.

[25] S. Loria, "TextBlob: Simplified Text Processing," [Online]. http://textblob.readthedocs.org/en/dev/authors.html. [Accessed 01 04 2016].

[26] F. A. Nielsen, "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs.," arXiv preprint arXiv:1103.2903, Technical University of Denmark, Lyngby, Denmark., 2011.

[27] F. Å. Nielsen, "A new ANEW: evaluation of a word list for sentiment analysis in microblogs," in *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages 718 in CEUR Workshop Proceedings*, Matthew Rowe, Milan Stankovic, Aba-Sah Dadzie, Mariann Hardey (editors), 2011 May, pp. 93-98.

[28] M. Thelwall, K. Buckley and G. Paltoglou, "Sentiment strength detection for the social Web," *Journal of the American Society for Information Science and Technology,* vol. 63, no. 1, pp. 163-173, Wiley Online Library, 2010.

[29] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai and A. Kappas, "Sentiment strength detection in short informal text," *Journal of the American Society for Information Science and Technology,* vol. 61, no. 12, p. 2544–2558, Wiley Online Library, 2010.