Business Analytics Report

Stock Market Forecasting based on Twitter mood

August 2017

Imperial College Business School

MSc Business Analytics
2016-2017
Words Count:

Abstract

“The stock market is a device for transferring money from the impatient to the patient” (Buffet, 2014). “The big money is not in the buying and the selling … but in the waiting” (Munger, 2014). Very few people, if not no one, could understand the stock market functionality better than Warren Buffet and Charlie Munger, two of the most successful investors in the world. Considering their words, one could ensure the theory that the stock market movement is nothing else than behavioural economics. Based on this concept, the present report investigates how collective mood can affect the stock market movement – Dow Jones Industrial Average returns change - on a daily basis. Daily tweets, which were collected by using both screen scraping and Application Programming Interfaces, were analysed with two distinct approaches; a unidimensional and a multidimensional approach. The former creates a unidimensional sentiment score measuring the positive vs negative public daily mood, whereas the latter creates and calculates 8 different daily mood dimensions - anger, anticipation, disgust, fear, joy, sadness, surprise and trust (Jockers, 2015). The dates of Brexit Election in United Kingdom and President Election in United States were used in order to test the validity of both sentiment approaches performance. A Stochastic Gradient Descent model was found as the optimal predictive model based on a technical analysis – prediction of DJIA returns based on its past lagged values – and then used, in combination with Twitter data, in order to accept or reject the theory that public collective mood can improve the stock market prediction performance. The results indicate that collective public mood can indeed enhance stock market prediction performance. The results indicate that collective public mood can indeed enhance stock market prediction results, indicating a substantial improvement of 771%.

Table of Contents

1 Introduction

The prediction of the future stock market returns constitutes a controversial issue that concerns not only the academical, but also the business world during the last decades (Han, 2012, Bollen, et al., 2010). The oldest and most established theory regarding the stock market, is based on the Efficient Market Hypothesis (EMH) (Bollen, et al., 2010). According to EMH, stock prices are a function of information i.e. news and, thus, all currently available information is already reflected in the current stock prices (Lo & MacKinlay, 2002). News is impossible to be predicted, which leads to Burton Malkiel's (1973) conclusion that stock prices are best characterized by a "random walk pattern". According to Malkiel's findings back to 1973, any change in the stock market prices which is due to a newly revealed information is random and cannot be predicted with more than a 50% accuracy. On the other side, competitors of this theory argue that there is a time interval between the reveal of a new information and the market response, during which predictions with higher than a 50% accuracy can be made (McKinley, 1999). Here again two different school of thoughts need to be taken into consideration; the fundamental and the technical analysis (Kalyani, et al., 2016). Fundamental analysis is based on a top-down approach starting from an economic analysis on a global level to a country, sector and company level respectively, examining the underlying factors that may affect a company's financial position (Kalyani, et al., 2016, Tomadaki-Balomenou, 2017). In simplest words, fundamental analysis refers to news data which act as economic or commercial indicators, whereas technical analysis is merely based on past stock returns (Kalyani, et al., 2016).

This research paper follows the fundamental analysis technique to investigate how the collective daily public mood can affect the stock market returns. In particular, the collective daily public mood is measured as the aggregated score of mood on a daily basis based on a daily's Twitter data sentiment analysis. Public mood acts as a proxy of people's reaction to news, as it has been proved by Kevin Bakhurst, the head of the BBC Newsroom, that more and more people share their emotions promptly to social media nowadays (Bakhurst, 2011), whereas Dow Jones Industrial Average Indexes closing prices act as the stock market's behaviour indicator (Yahoo Finance, 2017). Twitter (2017) was selected instead of other famous social media platforms such as Facebook mainly due to its less strict web-scraping limitations, but also to its more socioeconomic and politic nature. Public sentiment was selected instead of identifying 'bad' and 'good' news mainly because of the difficulty of acquiring tweets which refer only to news of specific companies or sectors i.e. it is impossible to distinguish between tweets which refer to the company Apple and to those which refer to the fruit apple, leading to completely irrelevant tweets that might distort the final prediction results. Nevertheless, Damasio (1994) has shown from early years how emotion and reason interact with each other to produce people's actions and decisions, meaning that collective public mood can influence the stock market in the same way as the news do (Bollen, et al., 2010).

The content of this report can be divided into four major parts. The first part refers to data acquisition. Two open data sources were particularly used in order to acquire the data: Twitter (Twitter, 2017) and Yahoo Finance (Yahoo Finance, 2017). The acquisition proccess includes both Application Programming Interfaces and web scraping techniques. The second part incorporates data pre-processing, including the handling of missing values and the formation of Twitter data using Natural Language Proccessing (NLP) techniques. The third part represents two sentiment analysis approaches; a unidimensional and a multidimensional approach. The unidimensional approach creates a daily positive or negative public mood score, whereas the multidimensional creates eight daily mood scores - anger, anticipation, disgust, fear, joy, sadness, surprise and trust (Jockers, 2015). Finally, the last part of the report, refers to predictive modelling. A technical analysis is done in order to find the optimal prediction model based on past values, and based on this model the effect of public mood on the model's prediction performance is examined.

2 Literature Review

Much progress has been noted during the last years regarding the prediction of the stock market behaviour (Han, 2012). Both technical and fundamental analysis have attracted a high amount of attention of several researchers and economists, leading to a substantial number of different studies and approaches. Moreover, the continuous development of advanced machine learning techniques has introduced a technological aspect to both kinds of analysis (Deboeck, 1994). The most recent of those studies have proved the existence of a strong correlation between sentiment analysis of news articles and social media posts regarding a specific company or the whole public world and the future trend of the specific company's stock or of the overall stock market, respectively (Kalyani, et al., 2016). Below follows a brief reference to past research on predicting stock market prices or movements based on a sentiment analysis of text data, focusing especially on social media data.

Fisher and Statman (2000) in their research studied the relationship between stock market returns and small, medium and large investor sentiment, respectively. Small and medium investor sentiment was measured based on weekly sentiment data of individuals and investment newsletter writers collected by the American Association of Individual Investors and Investors Intelligence Co respectively. Large investor sentiment was based on monthly sentiment data of Wall Street strategists compiled by Merrill Lynch & Company. A negative and statistically significant relationship between both small and large investor sentiment and S&P 500 returns was found, whereas it was concluded that medium investor sentiment does not have any impact on future stock market returns.

In another study which was done by Sisk (2013) and which is focused on a variety of social media data including Twitter, approximately 400 different features capturing people's mood were obtained from social media. A Principal Component Analysis was used in order to reduce the features obtained above from 400 to 30, capturing about 25% of the variance. An Ordinary Least Squares regression model was built based on these 30 features aiming to predict the stock behaviour. It was again concluded that news metadata can forecast short-term future stock price movements and volatility.

In another study, Bollen, Mao and Zeng (2011) examined the impact of daily collective mood states on the Dow Jones Industrial Average closing value changes of the following day. Daily tweets of a one-year period were obtained. Opinion Finder and Google-Profile of Mood States tools were then used in order to conduct a sentiment analysis of the tweets text on a daily basis, leading to a Self-Organizing Fuzzy Neural Network model with an accuracy of 87.6% in predicting the daily changes of the closing values of the DJIA and to a reduction of the Mean Average Percentage Error (MAPE) by more than 6%.

Californian researchers from University of California (2012) concluded that Twitter can forecast the stock market behaviour more accurately than any other investment strategy. According to the research team of the University of California, 150 companies in the S&P 500 Index were randomly selected. In contrast with the most studies that focus on tweets sentiment analysis, here tweets relevant to the specific companies were obtained, on an individual level, in order for the correlation between activity in Twitter – volume of tweets, retweets and links to other tweets or topics - and value of stock prices, as well as traded volume of the next day, to be examined. It was found that the number of trades was indeed correlated with the number of tweets, as well as with the number of connected components – the number of tweets including distinct topics regarding the same company - leading to a predictive model with an accuracy up to 11% more than other investment models. Stock prices were slightly correlated with connected components as well.

The strong relationship between investors' aggregated mood and future stock market returns was also proved by Siganos, Vagenas-Nanos and Verwijmeren (2014). In their study, the Facebook's Gross National Happiness Index (FGNHI) was used in order to measure the average positive or negative daily mood of people in each region. It was found that pessimistic sentiment is related to increases in trading volume and return volatility and vice versa.

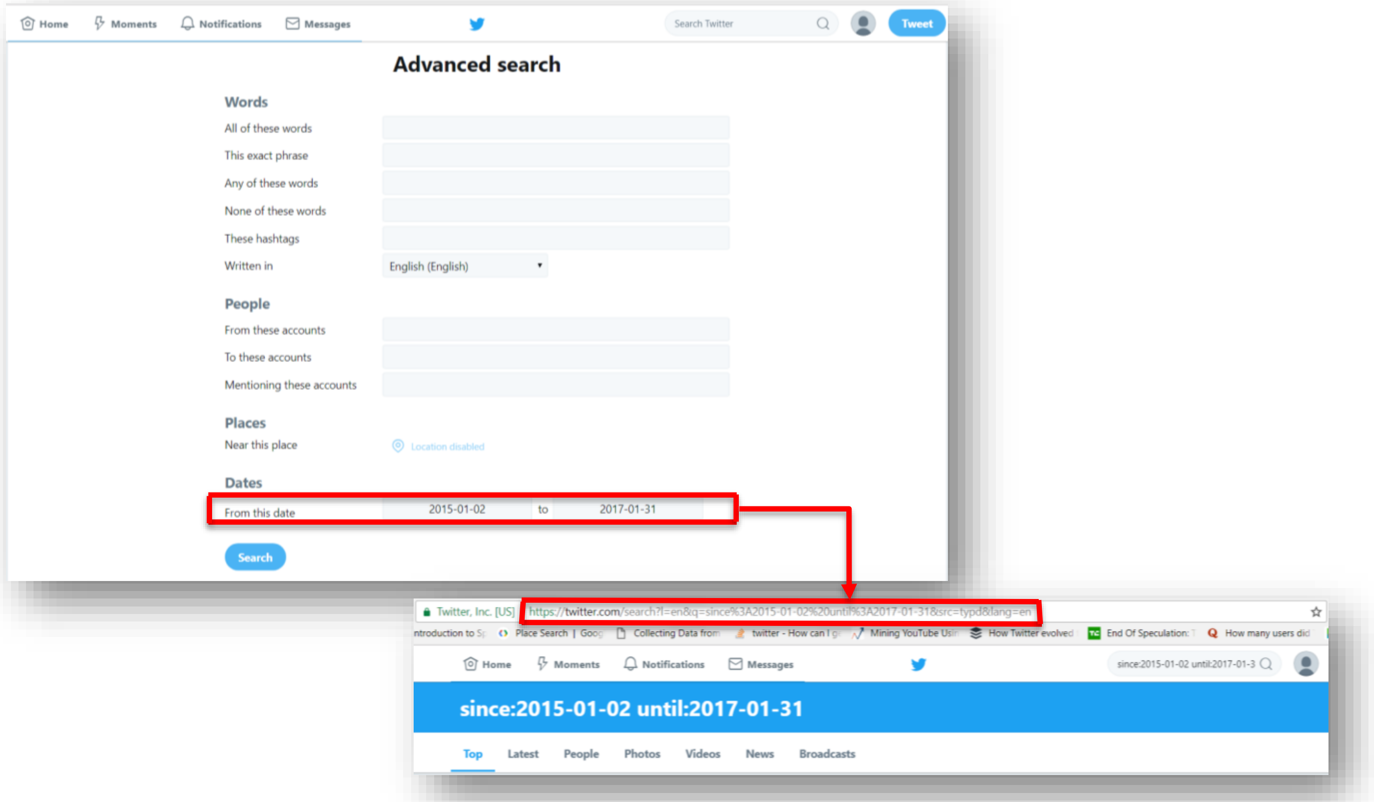
Economists of European Central Bank (2015) also reached the conclusion that Twitter can predict the stock market with a higher than 50% accuracy. Daily tweets containing the words "bullish" or "bearish" were used in order to measure the general mood of investors during each day. It was found that a one-point increase of the ECB's daily sentiment index led to a 12.56 points rise in DJIA returns of the following day, with a level of confidence of 99%. However, the bank noted that the Twitter mood indicator works only in the short-term and investors who wish to do long-term investments shouldn not rely on it.

3
Data Acquisition

Data acquisition refers to two main sources: Twitter (2017) and Yahoo Finance (2017).

3.1 Twitter

Daily tweets which were available to the public and able to capture users' general feelings and emotions during the time period from 2 January, 2015 to 31 January, 2017 were web scraped from Twitter. Due to Twitter's REST APIs limitations which permit the access to tweets of merely the last 7 days (Twitter, 2017), Python Selenium Web Driver Wrapper was used and the way that Twitter's advanced search engine works was adopted. An example of Twitter's advanced search engine functionality is presented in figure 1.



Advanced search

Words

All of these words

This exact phrase

Any of these words

None of these words

These hashtags

Written in: English (English)

People

From these accounts

To these accounts

Mentioning these accounts

Places

Near this place: Location disabled

Dates

From this date: 2015-01-02 to 2017-01-31

https://twitter.com/search?l=en&q=since%3A2015-01-02%20until%3A2017-01-31&src=typd&lang=en

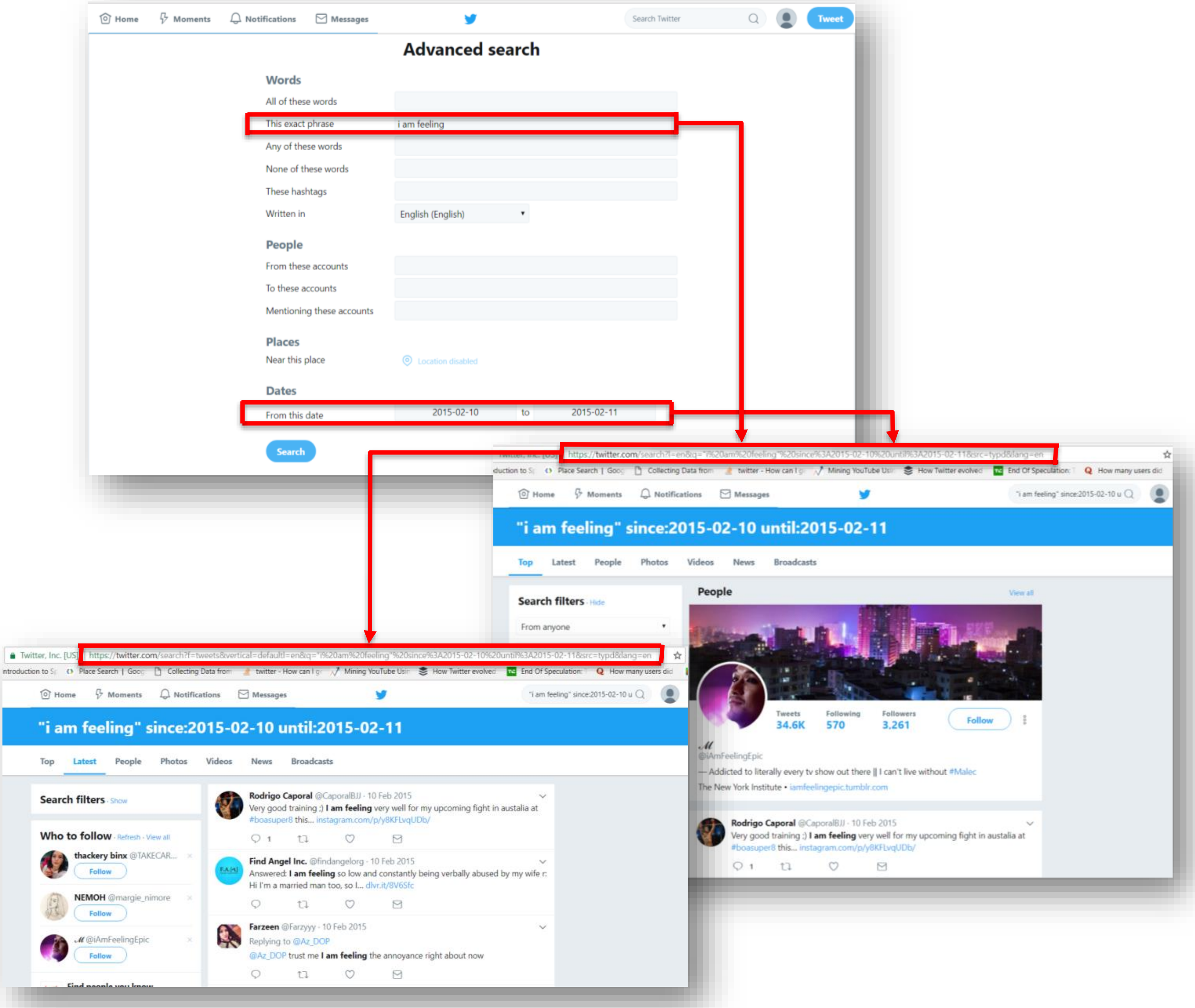
since:2015-01-02 until:2017-01-3

since:2015-01-02 until:2017-01-31

Figure 1. Twitter's advanced search engine; the top public tweets from 2 January, 2015 to 31 January, 2017 are

shown by this search option, which is equivalent to the following link :
https://twitter.com/search?l=en&q=since%3A2015-01-02%20until%3A2017-01-31&src=typd&lang=en

By using Python’s Selenium library, a browser was automatically opened up and the link of Twitter’s advanced search page was used in order to web scrape the tweets’ ids of each day. Furthermore, influenced by Bollen’s, Mao’s and Zeng’s work (2011), only tweets which contained the expressions: "I am", "I'm", "I feel", "I am feeling", "I'm feeling", "I don't feel" and "makes me" were considered to be public mood indicators and only those tweets were, therefore, web scraped, within the aforementioned time frame. This was again done by following Twitter’s advanced search engine mechanism, as shown in figure 2.



Advanced search

Words
All of these words
This exact phrase: i am feeling
Any of these words
None of these words
These hashtags
Written in: English (English)

People
From these accounts
To these accounts
Mentioning these accounts

Places
Near this place: Location disabled

Dates
From this date: 2015-02-10 to 2015-02-11

https://twitter.com/search?l=en&q="i%20am%20feeling"%20since%3A2015-02-10%20until%3A2015-02-11&src=typd&lang=en

"i am feeling" since:2015-02-10 until:2015-02-11

Search filters
From anyone

People
Tweets 34.6K · Following 570 · Followers 3,261
@iAmFeelingEpic
— Addicted to literally every tv show out there || I can't live without #Malec
The New York Institute • iamfeelingepic.tumblr.com

Twitter, Inc. [US] https://twitter.com/search?f=tweets&vertical=default&l=en&q="i%20am%20feeling"%20since%3A2015-02-10%20until%3A2015-02-11&src=typd&lang=en

"i am feeling" since:2015-02-10 until:2015-02-11

Who to follow
thackery binx @TAKECAR...
NEMOH @margie_nimore
@iAmFeelingEpic

Rodrigo Caporal @CaporalBJJ · 10 Feb 2015
Very good training ;) **I am feeling** very well for my upcoming fight in austalia at #boasuper8 this... instagram.com/p/y8KFLvqUDb/

Find Angel Inc. @findangelorg · 10 Feb 2015
Answered: **I am feeling** so low and constantly being verbally abused by my wife r: Hi I'm a married man too, so I... dlvr.it/8V6Sfc

Farzeen @Farzyyy · 10 Feb 2015
Replying to @Az_DOP
@Az_DOP trust me **I am feeling** the annoyance right about now

Figure 2. Twitter’s advanced seach engine; the top public tweets containing the word expression ‘I am feeling’ from

10 February, 2015 to 11 February, 2015 are shown by this search option, wjich is equivalent to the following link : https://twitter.com/search?l=en&q=%22i%20am%20feeling%22%20since%3A2015-02-10%20until%3A2015-02-11&src=typd&lang=en. By adding the quote ‘f=tweets&vertical=default&’ to the link (https://twitter.com/search?f=tweets&vertical=default&l=en&q="i%20am%20feeling"%20since%3A2015-02-10%20until%3A2015-02-11&src=typd&lang=en), it is defined that only tweets containing the phrase ‘I am feeling’ should be collected, avoiding Twitter accounts which contain the specific phrase just in their account name.

The specific process was followed for each day between the years 2015 and 2017, as well as for each one of the aforementioned word expressions.

After obtaining the tweets’ ids of each day within the aforementioned time frame, Twitter APIs were used in order to obtain the metadata – date of submission and text content - of each tweet id.

3.2 Yahoo Finance

The adjusted closing prices of the Dow Jones Industrial Average Index during the time period from 2 January, 2015 to 31 January, 2017 were downloaded from Yahoo Finance (2017).

4
Data Preprocessing

4.1 Missing Values

The dates were carefully selected in order to include a time period when Twitter had become famous enough and its users were active on a daily basis, leading to no missing values. However, DJIA's data were available only on weekdays, since the stock market does not operate on weekends. This means that DJIA's opening price on Monday will already incorporate the news that were released during the weekend. Thus, tweets which were created on the weekend were added to tweets which were created on Friday, in order to appropriately predict the stock market behaviour of Monday, leading to a dataset comprised of merely weekdays during the aforementioned time period. Moreover, the dates of bank holidays, when the stock market was closed, were found based on New York Time's Market Holiday Data web page (2017) and their tweets were similarly added to the previous day's tweets.

4.2 Twitter Data Transformation

Spam messages – tweets which started with the expression 'http:' or 'www:' (Bollen, et al., 2010) – were excluded from the dataset and all text was converted to lower case. Punctuation and stop-words – words, such as 'and', 'to', 'on', which are commonly used in the everyday life but do not convey any particular meaning – as well as words which started with an '@', indicating the name of the user, were removed from the dataset. Additional whitespaces between words created by the exclusion of punctuation were also removed. Repeating letters – more than two in a row - were replaced by the letter itself i.e. 'happyyyy' was replaced by 'happy' and spelling mistakes were adjusted based on Peter Norving's Spelling Corrector mechanism (Norvig, 2016). Furthermore, it was noted that Twitter's quotation marks were not treated as punctuation, but as stings with the consequence of many words starting with or ending in quotation marks. Therefore, another limitation, indicating that each word should start with or end in an alphabet letter, was introduced, leading to a dataset of mere text. Finally, all tweets referring to the same dates were grouped to a single string, which was then split into a list of words. A representation of a Twitter data tranformation example is shown in figure 3, whereas a small sample of the final dataset of text is presented in figure 4.

Twitter Data Tranformation Process

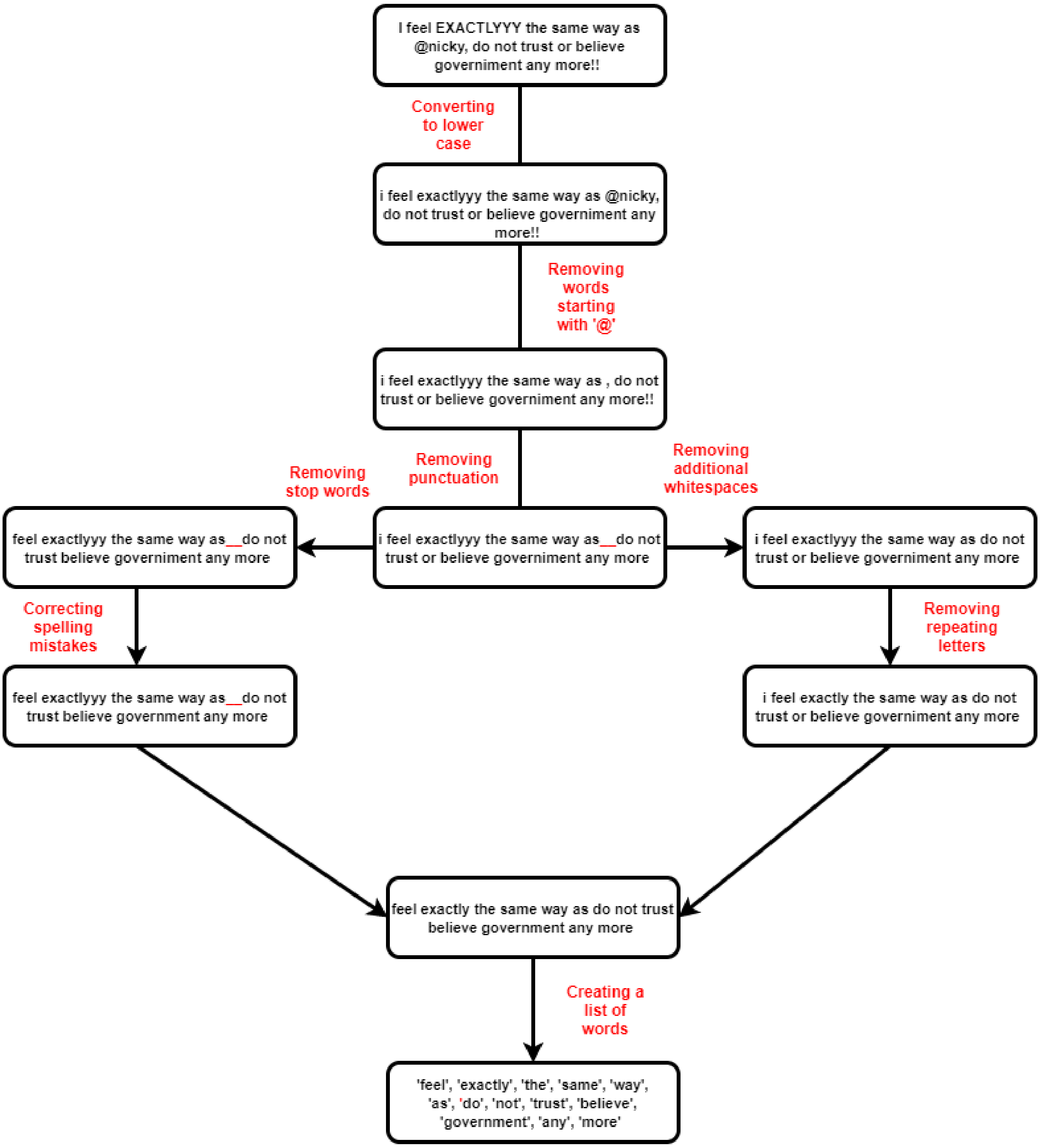
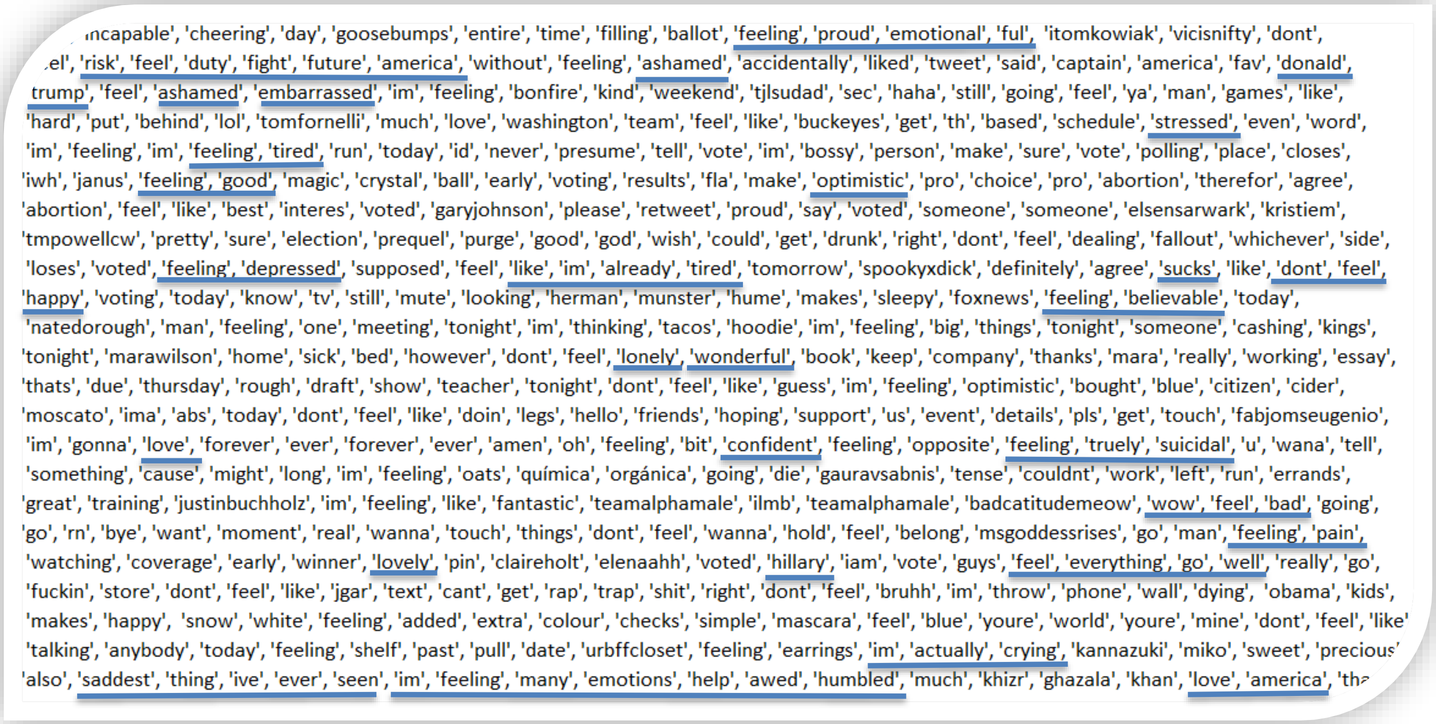

I feel EXACTLYYY the same way as @nicky, do not trust or believe government any more!!
Converting to lower case
i feel exactlyyy the same way as @nicky, do not trust or believe government any more!!
Removing words starting with '@'
i feel exactlyyy the same way as , do not trust or believe government any more!!
Removing stop words
Removing punctuation
Removing additional whitespaces
feel exactlyyy the same way as__do not trust believe government any more
i feel exactlyyy the same way as__do not trust or believe government any more
i feel exactlyyy the same way as do not trust or believe government any more
Correcting spelling mistakes
Removing repeating letters
feel exactlyyy the same way as__do not trust believe government any more
i feel exactly the same way as do not trust or believe government any more
feel exactly the same way as do not trust believe government any more
Creating a list of words
'feel', 'exactly', 'the', 'same', 'way', 'as', 'do', 'not', 'trust', 'believe', 'government', 'any', 'more'

Figure 3. An example of the standard Twitter data tranformation process

Text of tweets on 8 November, 2017



```
incapable', 'cheering', 'day', 'goosebumps', 'entire', 'time', 'filling', 'ballot', 'feeling', 'proud', 'emotional', 'ful',  'itomkowiak', 'vicisnifty', 'dont',
el', 'risk', 'feel', 'duty', 'fight', 'future', 'america', 'without', 'feeling', 'ashamed', 'accidentally', 'liked', 'tweet', 'said', 'captain', 'america', 'fav', 'donald',
trump', 'feel', 'ashamed', 'embarrassed', 'im', 'feeling', 'bonfire', 'kind', 'weekend', 'tjlsudad', 'sec', 'haha', 'still', 'going', 'feel', 'ya', 'man', 'games', 'like',
'hard', 'put', 'behind', 'lol', 'tomfornelli', 'much', 'love', 'washington', 'team', 'feel', 'like', 'buckeyes', 'get', 'th', 'based', 'schedule', 'stressed', 'even', 'word',
'im', 'feeling', 'im', 'feeling', 'tired', 'run', 'today', 'id', 'never', 'presume', 'tell', 'vote', 'im', 'bossy', 'person', 'make', 'sure', 'vote', 'polling', 'place', 'closes',
'iwh', 'janus', 'feeling', 'good', 'magic', 'crystal', 'ball', 'early', 'voting', 'results', 'fla', 'make', 'optimistic', 'pro', 'choice', 'pro', 'abortion', 'therefor', 'agree',
'abortion', 'feel', 'like', 'best', 'interes', 'voted', 'garyjohnson', 'please', 'retweet', 'proud', 'say', 'voted', 'someone', 'someone', 'elsensarwark', 'kristiem',
'tmpowellcw', 'pretty', 'sure', 'election', 'prequel', 'purge', 'good', 'god', 'wish', 'could', 'get', 'drunk', 'right', 'dont', 'feel', 'dealing', 'fallout', 'whichever', 'side',
'loses', 'voted', 'feeling', 'depressed', 'supposed', 'feel', 'like', 'im', 'already', 'tired', 'tomorrow', 'spookyxdick', 'definitely', 'agree', 'sucks', 'like', 'dont', 'feel',
'happy', 'voting', 'today', 'know', 'tv', 'still', 'mute', 'looking', 'herman', 'munster', 'hume', 'makes', 'sleepy', 'foxnews', 'feeling', 'believable', 'today',
'natedorough', 'man', 'feeling', 'one', 'meeting', 'tonight', 'im', 'thinking', 'tacos', 'hoodie', 'im', 'feeling', 'big', 'things', 'tonight', 'someone', 'cashing', 'kings',
'tonight', 'marawilson', 'home', 'sick', 'bed', 'however', 'dont', 'feel', 'lonely', 'wonderful', 'book', 'keep', 'company', 'thanks', 'mara', 'really', 'working', 'essay',
'thats', 'due', 'thursday', 'rough', 'draft', 'show', 'teacher', 'tonight', 'dont', 'feel', 'like', 'guess', 'im', 'feeling', 'optimistic', 'bought', 'blue', 'citizen', 'cider',
'moscato', 'ima', 'abs', 'today', 'dont', 'feel', 'like', 'doin', 'legs', 'hello', 'friends', 'hoping', 'support', 'us', 'event', 'details', 'pls', 'get', 'touch', 'fabjomseugenio',
'im', 'gonna', 'love', 'forever', 'ever', 'forever', 'ever', 'amen', 'oh', 'feeling', 'bit', 'confident', 'feeling', 'opposite', 'feeling', 'truely', 'suicidal', 'u', 'wana', 'tell',
'something', 'cause', 'might', 'long', 'im', 'feeling', 'oats', 'química', 'orgánica', 'going', 'die', 'gauravsabnis', 'tense', 'couldnt', 'work', 'left', 'run', 'errands',
'great', 'training', 'justinbuchholz', 'im', 'feeling', 'like', 'fantastic', 'teamalphamale', 'ilmb', 'teamalphamale', 'badcatitudemeow', 'wow', 'feel', 'bad', 'going',
'go', 'rn', 'bye', 'want', 'moment', 'real', 'wanna', 'touch', 'things', 'dont', 'feel', 'wanna', 'hold', 'feel', 'belong', 'msgoddessrises', 'go', 'man', 'feeling', 'pain',
'watching', 'coverage', 'early', 'winner', 'lovely', 'pin', 'claireholt', 'elenaahh', 'voted', 'hillary', 'iam', 'vote', 'guys', 'feel', 'everything', 'go', 'well', 'really', 'go',
'fuckin', 'store', 'dont', 'feel', 'like', 'jgar', 'text', 'cant', 'get', 'rap', 'trap', 'shit', 'right', 'dont', 'feel', 'bruhh', 'im', 'throw', 'phone', 'wall', 'dying',  'obama', 'kids',
'makes', 'happy',  'snow', 'white', 'feeling', 'added', 'extra', 'colour', 'checks', 'simple', 'mascara', 'feel', 'blue', 'youre', 'world', 'youre', 'mine', 'dont', 'feel', 'like'
'talking', 'anybody', 'today', 'feeling', 'shelf', 'past', 'pull', 'date', 'urbffcloset', 'feeling', 'earrings', 'im', 'actually', 'crying', 'kannazuki', 'miko', 'sweet', 'precious
'also', 'saddest', 'thing', 'ive', 'ever', 'seen', 'im', 'feeling', 'many', 'emotions', 'help', 'awed', 'humbled', 'much', 'khizr', 'ghazala', 'khan', 'love', 'america', 'tha
```

Figure 4. A small sample of the exact text which is related to the tweets on 8 November, 2017 – the day when Donald Trump was elected as the US new president

5

Sentiment Analysis

5.1 Unidimensional Sentiment Analysis

The traditional approach described in Peter Turney's research paper (2002) was used for the sentiment analysis in this report. Although there is much progress on the specific subject since then, the specific approach was selected due to its simplicity; it does not require any labelled data for training, as it constitutes an unsupervised technique (Bonzanini, 2015). Acquiring from online sources a large-scale and valid dataset of tweets, labelled by positivity or negativity, to be used in testing, would be really difficult if not impossible.

According to Turney's approach, the Semantic Orientation of a term is defined as 'the difference between its associations with positive and negative words' (Bonzanini, 2015), as it is shown in equation (1). Here it should be noted that the aforementioned association of each word is calculated against a set of positive and negative terms (Bonzanini, 2015) based on the online dictionary created for sentiment analysis by Minqing Hu and Bing Liu (2004). The semantic orientation of a word is positive if the word is often associated with words from the positive lexicon, whereas negative if the word is often associated with words from the negative lexicon. For neural words the semantic orientation equals zero (Bonzanini, 2015).

$$SO(t) = \sum_{t' \in V+} PMI(t,t') - \sum_{t' \in V-} PMI(t,t') \quad (1)$$

where $\sum_{t' \in V+} PMI(t,t')$: the sum of the 'closeness' score between the term t and the term t' which is included in the dictionary of positive words, $\sum_{t' \in V-} PMI(t,t')$: the sum of the 'closeness' score between the term t and the term t' which is included in the dictionary of negative words (Bonzanini, 2015).

Each word's closeness to a positive or negative word is measured by using the Pointwise Mutual Information (PMI) measure (Bonzanini, 2015). According to PMI, the closeness between two points is identified as 'the discrepancy between the probability of their coincidence given their joint distribution and their individual distributions, assuming independence.' ('Pointwise mutual information', 2017). The probability of observing the term t can be calculated as the Document Frequency of the term t divided by the absolute number of documents D (Bonzanini, 2015). Each document D is represented by a tweet; hence the Document Frequency can be interpreted as the total number of documents or tweets where the term t occurs. Co-occurrent terms follow the same logic. Mathematically shown:

$$PMI(t_1, t_2) = \log\frac{P(t_1 \cap t_2)}{P(t_1)P(t_2)} \quad (2)$$

$$P(t) = \frac{DF_{(t)}}{|D|} \quad (3)$$

$$P(t_1 \cap t_2) = \frac{DF_{(t_1 \cap t_2)}}{|D|} \quad (4)$$

where t_1, t_2: terms, P(t): the probability of observing the term t, $P(t_1 \cap t_2)$: the probability of observing the terms t1 and t2 occurring together, $DF_{(t)}$: document frequency of the term t, $DF_{(t_1 \cap t_2)}$: document frequency of the co-occurrent terms t_1 and t_2, D: documents or tweets (Bonzanini, 2015).

After obtaining the Semantic Orientation of each word, the final sentiment score of each tweet was defined as the aggregated Semantic Orientation score of its words. Here, it should be noted that terms, which followed the word 'not' or 'dont', were identified and their sign was inversed. The calculated Semantic Orientation of the words 'not' and 'dont' was then removed from the tweet's final sentiment score. Finally, the sentiment score, and thus the public mood, of each day was defined as the average semantic score of tweets which referred to that day.

A representation of the final daily sentiment index can be seen in both figures 3 and 4. Figure 3 depicts the actual movement of DJIA returns in red and the sentiment index change in blue on a daily basis. As it will be explained in chapter 6, the DJIA return prices on day t reflect the mood values of the past 3 days (t – 3). Figure 4 represents the daily emotional varianve through time from 2 January, 2015 to 31 January, 2017. The distinct blue data point represents the public mood on 23 June, 2016, when Brexit was elected in United Kingdom, whereas the green data point refers to the value of the sentiment index on 8 November, 2016, when Donald Trump was elected as the United States new president.

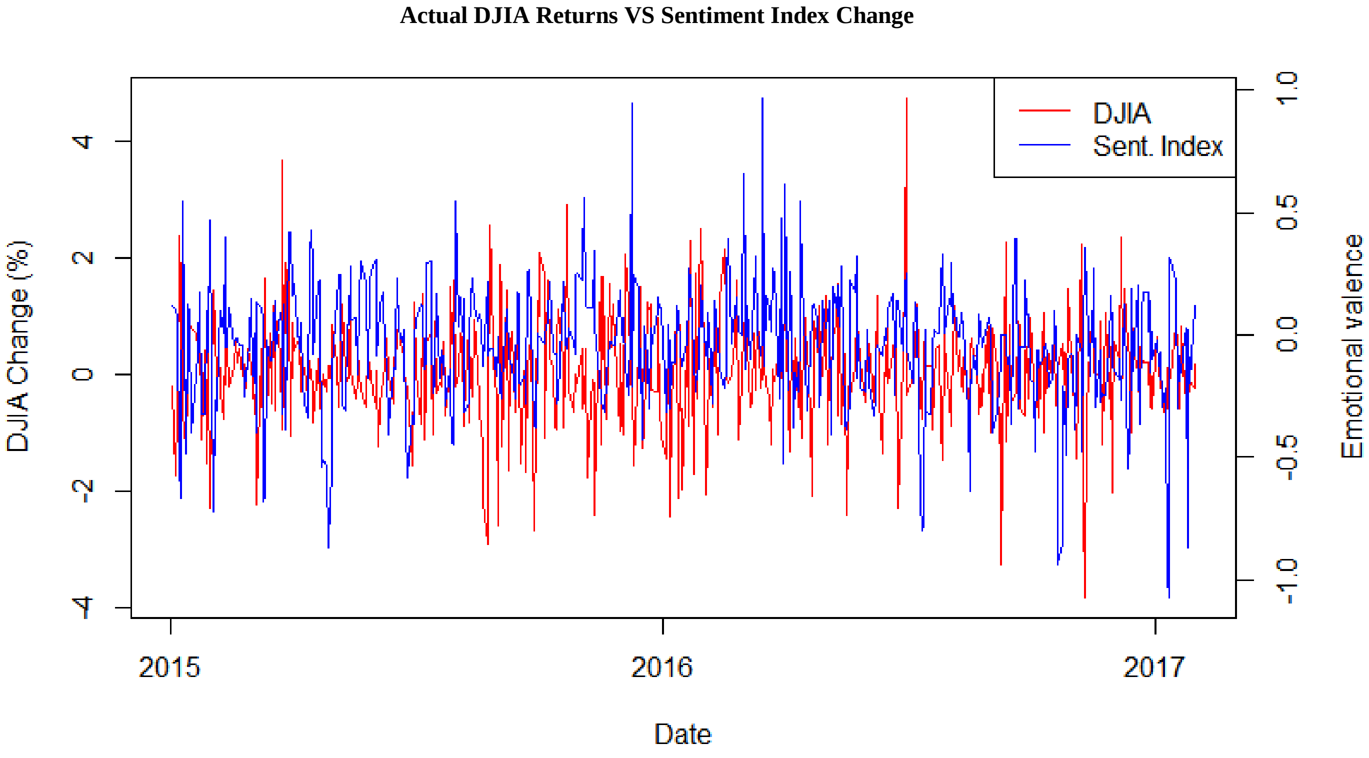


Figure 3. DJIA actual change compared to sentiment indexes change between the years 2015 and 2017

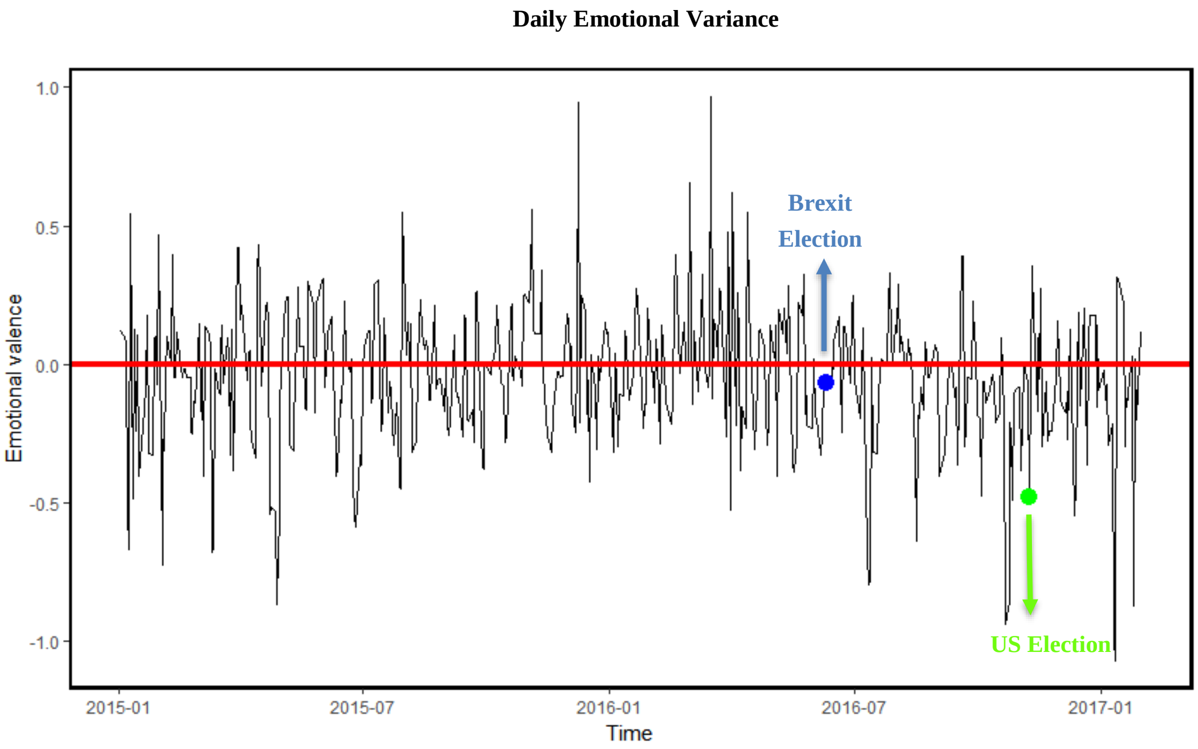


Figure 4. Daily emotional variance from 2 January, 2015 to 31 January, 2017

5.2 Multidimensional Sentiment Analysis

When using a unidimensional sentiment score to capture public mood, one should note that in this way the rich and multidimensional nature of human mood is ignored (Bollen, et al., 2010). For example, a tweet can express fear and joy at the same time i.e. a student who is ready to undertake his first substantial project. To capture additional mood aspects Matthew L. Jocker's new package, so called Syuzhet, which identifies eight different mood dimensions - anger, anticipation, disgust, fear, joy, sadness, surprise and trust - based on four different opinion lexicons was used (Jockers, 2015). In this way, a score for each word for each of the eight dimensions was created and the aggregated weighted score of the words, which referred to tweets of the same date, were defined as the final daily score for each of the eight mood dimensions. Figure 5 shows the frequency of each of the 8 sentiments within the whole dataset of tweets, even when combined with other sentiments. Table 1 represents frequency in numbers.

Frequency of 8 Sentiments

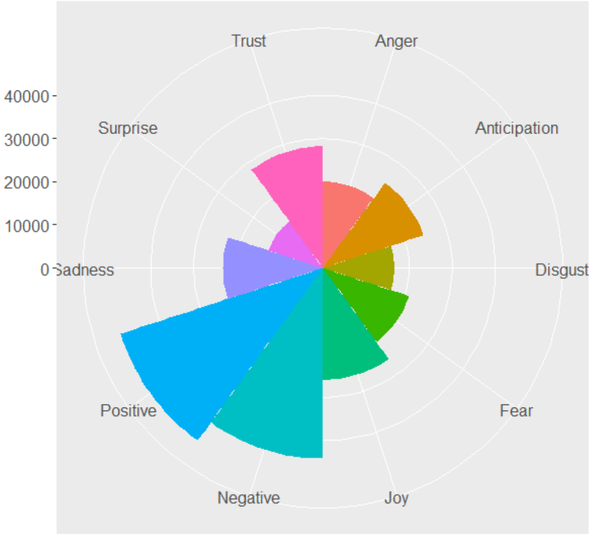


Figure 5. Frequency of 8 sentiments; sadness, surprise, trust, anger, anticipation, disgust, fear and joy

Frequency of 8 Sentiments

	Frequency
Anger	20226
Anticipation	24726
Disgust	16773
Fear	21310
Joy	26179
Sadness	23421
Surprise	13387
Trust	28441
Negaive	44715
Positive	49955

Table 1. Frequency of 8 sentiments in numbers

The Brexit and US Election events were again used in order to examine the validity of the occured sentiment index. Concidering the fact that the sentiment indexes value is positive – high value of trust and joy – during both events, it can be inferred that the multidimensional option fails to accurately predict the public mood. However, on the date of US President Election a considerably high, compared to others, aggregated sentiment score can be observed, meaning that the specific approach lacks mainly in seperating the sentiments into the correct mood states. Figure 6 captures the change of the 8 mood states between the years 2015 and 2017. Again the blue and green arrows refer to the date of the Brexit and US Election events, respectively. Data points could not be used in this occasion, as the figure represents 8 different mood states per day.

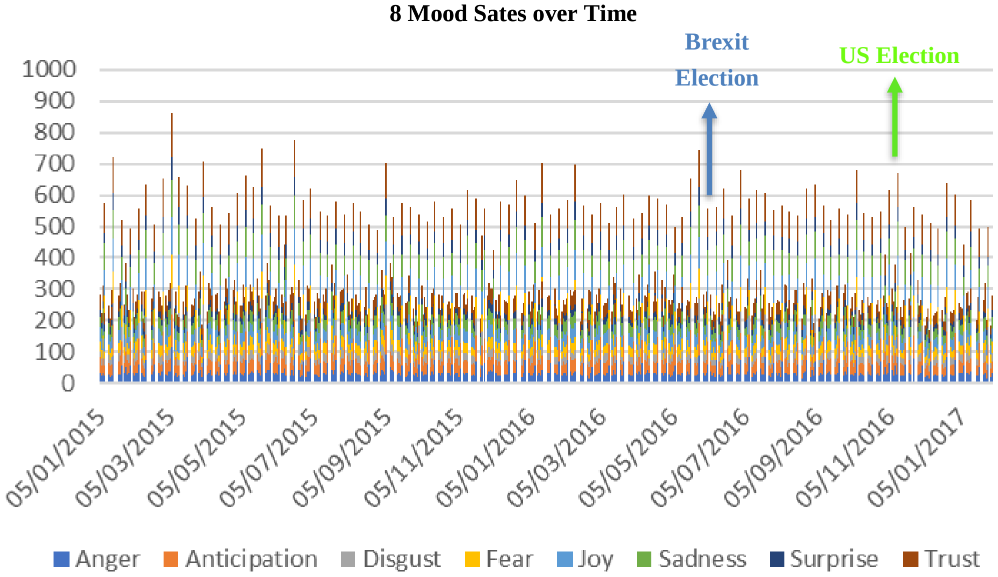


Figure 6. Change of 8 mood states over time from 2 January, 2015 to 31 January, 2017

6

Predictive Modelling

DJIA’s return prices were calculated based on the well-known formula (5), constituting the values to be predicted.

$$DJIA\ returns = \frac{Adj.Closing\ Price_t - Adj.Closing\ Price_{t-1}}{Adj.Closing\ Price_{t-1}} \quad (5)$$

Initially, a prediction based on a technical analysis was done in order to find the optimal prediction model based on past return prices. The same model, but based on public mood, was then used in order to accept or reject the theory that collective public mood can predict more accurately the stock market. Here it should be noted that the scope of this report is to just prove that twitter data can improve the stock market prediction performance, and not to find the optimal prediction model based on twitter data.

6.1 Technical Analysis: DJIA Returns Prediction based on Past Return Prices

6.1.1 Predictive Modelling based on Past Values

The Historical Moving Average constitutes the simplest and most common technique used for financial forecasting in the business world today. Hence, in this report the Root Mean Squared Error of Historical Moving Average was first calculated and then used as a base to obtain the out of sample residual squared compared to different machine learning techniques. The model with the best performance – greater Out Of Sample Residual squared - was considered to be the optimal model based on past DJIA returns. Several different machine learning techniques – Linear, Lasso, Ridge and Elastic Net Regression, Random Forest, Stochastic Gradient Descent and Support Vector Machines - were examined. One, two, three, four, seven, fifteen, thirty and sixty days lagged variables were tested within a 5-fold cross-validated linear regression model. Three days lagged variable gave the smaller residual and therefore constituted the variable that was used in Random Forest, SVM and SGD. One and thirty days lagged variables came second and third in residuals, thus all of them, including three days lagged variable, were used in a single Lasso, Ridge and Elastic Net Regression model, since their penalization functionality will shrink to zero unnecessary coefficients. Oil and Gold past prices were also used as regressors in the three aforementioned models as macroeconomic indicators. Here it should be noted that all models' hyperparameters were tuned within a 5-fold cross-validated grid search and boosted by an ADA Boost algorithm.

However, the fact of instability was obvious and affected the prediction results. Instability is considered to be one of the most critical issues in macroeconomic and financial forecasting (Goyal & Welch, 2003). 'To handle such instability, it is quite common to use only the most recent observations to estimate parameters of forecasting models rather than all available observations, the so-called "rolling estimation" method.' (Inoue, et al., 2017). For example, if there is a high demand for a product at 11, chances are that there will be some demand at 11.01 as well. This is the so-called autocorrelation fact, indicating that the outcome of a prediction is not only affected by X_t, but also by X_{t-1}, X_{t-2} and so on. Hence, a rolling window of 20 days was introduced and used in combination with each of the aforementioned machine learning techniques. The optimal window size was identified based on a 5-fold cross validated linear regression model.

6.1.2 Prediction Results

The results of the DJIA returns forecasting based on historical data are shown in table 2. Stochastic Gradient Descent was the only machine learning technique which outperformed, or performed slightly better than, the Historical Moving Average.

Out Of Sample Residual Squared Error

	Linear Regression Lagged 1 day	Linear Regression Lagged 3 days	Linear Regression Lagged 30 days	Lasso Regression	Ridge Regression	Elastic Net Regression	Random Forest	SVR	SGD
RsqOOS	-0.000547449	-0.000204677	-0.000240607	-0.0001293	-0.002199	-0.0001293	-0.00072	-0.00064	0.000187

Table 2. Calculating the Out Of Sample Residual Squared between the Moving Historical Average and other Machine Learning techniques.

6.2 Fundamental Analysis : DJIA Returns Prediction based on Collective Public Mood

Since the Stochastic Gradient Descent was the only technique which outperformed the Historical Moving Average based on past returns, the further analysis based on Twitter data was focused only on that model. As it has been already mentioned the sentiment analysis was done on two levels; a unidimensional and a multidimensional level. Hence, two different predictive models were created based on Twitter data.

6.2.1 DJIA Returns Forecasting based on Unidimensional Mood Data

Three days lagged DJIA past returns along with three days lagged sentiment index values were used within a 5-fold cross validated Stochstic Gradient Descent model. The model's hyperparameters were again tuned within a 20 days rolling window and boosted by an ADA Boost algorithm. Table 3 depicts the model's prediction results.

Out Of Sample Residual Squared Error

	SGD based on past prices	SGD based on past prices & unidimensional sentiment index
RsqOOS	0.000186818	0.001628962

Table 3. Calculating the Out Of Sample Residual Squared between SGD obtained from a technical analysis to a SGD obtained from a fundamental analysis (based on the unidimensional approach)

6.2.2 DJIA Returns Forecasting based on Multidimensional Mood Data

Exactly the same procedure - as in chapter 6.2.1 - was followed, with the excpetion that the three days lagged values of the 8 distinct mood states were used as regressors instead of the three days lagged sentiment index values. Table 4 shows the model's prediction results.

Out Of Sample Residual Squared Error

	SGD based on past prices	SGD based on past prices & multidimensional sentiment index
RsqOOS	0.0001868	0.000193432

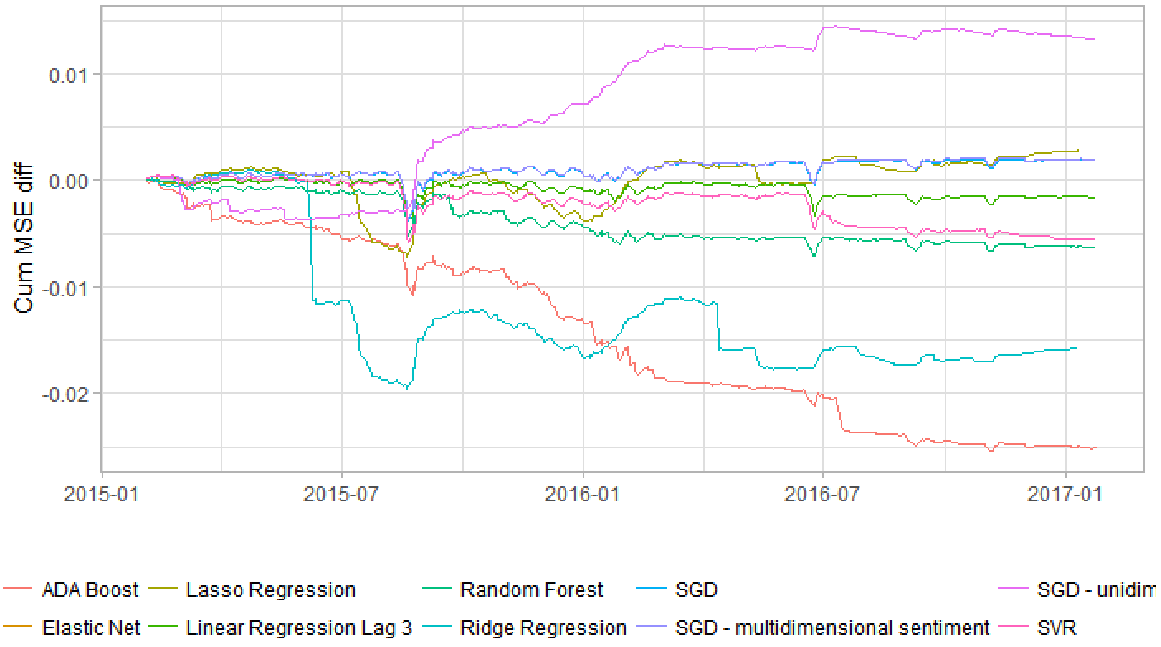
Table 4. Calculating the Out Of Sample Residual Squared between SGD obtained from a technical analysis to a SGD obtained from a fundamental analysis (based on the multidimensional approach)

6.2.1 Comparison

Considering the Out Of Sample Residual Squared between the Stochastic Gradient Descent models of technical and fundamental analysis, it was proved that both fundamental approaches can imrove significantly the stock market prediction performance. Multidimensional approach noted a slight increase of 0.00019 points, whereas unidimensional approach succeeded a substantial rise of 0.0016 points. Figure 7 represents the performance -

measured in Cumulatuve Mean Square Error Differential from Historical Moving Average values - of all the machine learning techniques which were tried for the scope of this report.Someone could easily notice how Stochastic Gradient Descent model outperformes all other techniques. A small gap in the early August, 2015 which couldn't be predicted from any of the aforementioned techniques could raise interesting questions and motive for further investigation in the future.

Cumulative MSE differential – Moving Average vs ML Techniques



7 Conclusion

The analysis mentioned in this report proves that indeed Twitter data can have a significant positive impact on the stock market behavior prediction performance. The proposed model, a Stochastic Gradient Descent model based on a combination of past stock returns and a unidimensional sentiment index, outperformed the optimal – based on the current analysis – machine learning technique based on technical data by 771%.

The current report could be used as a base for further research on the topic. Sentiment analysis is a sector which is still developing and further, more advanced sentiment techniques are anticipated to appear in the future. Moreover, this report proves the power of Twitter data, however it lacks of accurate and targeted data, as it will be explained in chapter 8. Precise data obtained nt only from Twitter but from other online sources as well could have substantial impact on stock market forecasting.

8 Limitations

One of the greatest limitations with which the current report is dealing is the quality of Twitter data. Mainly due to privacy reasons, Twitter APIs permit access to Twitter data only for the last 7 days. However, recognizing the continuous increasing demand for Twitter data by both tech companies and researchers, Twitter has started cooperating with third parties who manipulate, based on each customer's needs, and sell its data. Therefore, acquiring those data by screen scraping past Twitter pages does not seem to be the most effective way. When such patents exist, it becomes apparent that even web scraping will face privacy obstacles. Indeed, the number of tweets which is acquired per day is considerably lower than what is anticipated, raising the question of whether those data could be representative of all users' tweets. Furthermore, due to the enormous number of daily tweets and hastags by each kind of user, it seems impossible to focus on specific topics without involving a considerably high amount of irrelevant text. Using a dataset of tweets which refer to a specific topic i.e. Apple company or sector i.e. tech and which were created from people who are interested in financial and politic matters, would definitely improve prediction results.

Furthermore, when it comes to sentiment analysis, another limitation refers to the manipulation of negative words. Identifying negative words which completely inverse the meaning of a phrase constitutes a matter which has raised a lot of attention lately (Liu, 2015). For example, the phrase 'I do not trust any government any more' declares a negative mood by the user who wrote it, however if someone looks at distinct words the word 'trust' would get a high positive score penalizing the negative score of the word 'not', leading to a positive mood identification. In this report, the negative words 'don't' and 'not' were found and the sign of the next word's score was inversed. However, there are a lot of cases where again the inverse meaning of a word cannot be captured. For instance, in the phrase 'not so bad', the word which follows the negative word 'not' is 'so', which is a completely neutral word and its score will be 0. Then the score of 'so' will be inversed – remain zero – whereas the high negative score of the word 'bad' will lead to the faulty conclusion of a negative mood identification. A solution to this problem could be to identify the first noun after such negative words and inverse their sign – library NLTK in Python is able to identify which words are nouns from a list of words.

More advancements in this problem are anticipated soon, as it is still open among sentiment researchers.

9
References

Bakhurst, K., 2011. *How has social media changed the way newsrooms work?,* United Kingdom: BBC News, The Editors.

Bing, L. & Minqing, H., 2004. *Mining and Summarizing Customer Reviews,* Seattle: ACM SIGKDD International Conference on Knowledge .

Bollen, J., Mao, H. & Zeng, X.-Z., 2010. *Twitter Mood predicts the Stock Market,* United States: Cornell University Library.

Bonzanini, M., 2015. *'Mining Twitter Data with Python (Part 6 – Sentiment Analysis Basics)'.* [Online] Available at: https://marcobonzanini.com/2015/05/17/mining-twitter-data-with-python-part-6-sentiment-analysis-basics/
[Accessed 8 August 2017].

Damasio, A., 1994. *Descartes' Error: Emotion, Reason, and the Human Brain.* 1st ed. New York: G. P. Putnam's Sons.

Deboeck, G. J., 1994. *Trading on the Edge: Neural, Genetic, and Fuzzy Systems for Chaotic Financial Markets.* 1st ed. United States: John Wiley & Sons, Inc..

Duggan, W., 2014. *Is Warren Buffett Secretly Taking Your Money?.* [Online] Available at: https://www.fool.com/investing/general/2014/05/07/is-warren-buffett-secretly-taking-your-money.aspx [Accessed 30 August 2017].

Fisher, K. L. & Statman, M., 2000. Investor Sentiment and Stock Returns. *Financial Analysts Journal,* 56(2).

Goyal, A. & Welch, I., 2003. *Predicting the Equity Premium with Dividend Ratios,* United States: Management Science.

Han, Z., 2012. *Data and Text Mining of Financial Markets using News and Social Media,* Manchester: The University of Manchester.

Inoue, A., Jin, L. & Rossi, B., 2017. Rolling Window Selection for Out-of-Sample Forecasting with Time-Varying Parameters. *Journal of Econometrics,* 196(1).

Jockers, M. L., 2015. *Revealing Sentiment and Plot Arcs with the Syuzhet Package.* [Online] Available at: http://www.matthewjockers.net/2015/02/02/syuzhet/
[Accessed 28 August 2017].

Kalyani, J., Bharathi, H. N. & Jhothi, R., 2016. *STOCK TREND PREDICTION USING NEWS SENTIMENT ANALYSIS,* United States: Cornell University.

Liu, B., 2015. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Toronto: Morgan & Claypool Publishers.

Lo, A. & MacKinlay, A. C., 2002. *A Non-Random Walk Down Wall Street.* 5th ed. United States: Princeton University Press.

Malkiel, B., 1973. *A Random Walk Down Wall Street.* 1st ed. United States: W. W. Norton & Company, Inc..

McKinley, K., 1999. Stock market efficiency and insider trading. *Political Economy,* Volume 8 July 1999.

Norvig, P., 2016. *How to Write a Spelling Corrector.* [Online] Available at: http://norvig.com/spell-correct.html[Accessed 17 August 2017].

'Pointwise mutual information', 2017. *Wikipedia: The Free Encyclopedia.* [Online] Available at: https://en.wikipedia.org/wiki/Pointwise_mutual_information [Accessed 20 August 2017].

Siganos, A., Vagenas-Nanos, E. & Verwijmeren, P., 2014. Facebook's daily sentiment and international stock markets. *Journal of Economic Behavior & Organization,* Volume 107.

Sisk, J., 2013. *Methods and systems for predicting market behavior based on news and sentiment analysis,* United States: 20130138577 A1.

The New York Times, 2017. *Business Day Markets.* [Online] Available at: http://markets.on.nytimes.com/research/markets/holidays/holidays.asp?display=market&exchange=NSQ [Accessed 20 August 2017].

Tomadaki-Balomenou, A., 2017. *Predicting Industry Stock Returns,* London: Imperial College Business School.

Turney, P. D., 2002. *Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews,* Ontario: National Research Council of Canada.

Twitter, 2017. *Twitter.* [Online] Available at: https://twitter.com/ [Accessed 10 August 2017].

Waugh, R., 2012. The Tweets ARE paved with gold: Twitter 'predicts' stock prices more accurately than any investment tactic, say scientists. *Daily mail*, 26 March.

Williams-Grut, O., 2015. The ECB says Twitter can predict the stock market. *Business Insider UK*, 22 July.

Yahoo Finance, 2017. *YAHOO! FINANCE.* [Online] Available at: https://finance.yahoo.com/quote/%5EDJI?p=%5EDJI [Accessed 20 August 2017].