SOCIAL MEDIA INTERACTIONS PREDICT THE STOCK MARKET

Updated April 2019
I, Marouan Jazouli, affirm that I applied ethics to the design process and in the selection of the final proposed design. And that, the designer has held the safety of the public to be paramount and has addressed this in the presented design wherever may be applicable.

Marouan Jazouli

Approved by the Supervisor(s)

Dr. T. Rachidi
Acknowledgements

It is a matter of great pleasure to present this final report on “Social Media Interactions Predict the Stock market”.

The success of this project would not have been possible without the constant help, patience and advice of my supervisor Dr. Tajje-eddine Rachidi. Dr. Tajje-eddine Rachidi helped choose the right project given my expertise, skills and future career and I am very thankful for it. If it was not for his expertise and counsel, this project would not have seen the development it reaches today. Thanks to his openness to discussion and knowledge, I was able to acquire numerous notions that I know will help me throughout my professional career.

I would also like to take a moment to address my sincerest gratitude to each and every one that has been here for me for the last couple of years. I would not have made it this far if it was not for my loving parents, my amazing friends and my dedicated professors who have supported me and believed in me.
Table of Contents

Acknowledgements .......................................................................................................................... ii

List of Figures ................................................................................................................................. v

Abstract ........................................................................................................................................ vi

Introduction ..................................................................................................................................... 1

Project Scope ................................................................................................................................... 2

4.1 Project Overview ...................................................................................................................... 2

4.2 Target Audience ..................................................................................................................... 2

4.3 Objectives .............................................................................................................................. 2

  4.3.1 Project Objectives ............................................................................................................. 2

  4.3.2 Learning Objectives ......................................................................................................... 3

STEEPLE Analysis ........................................................................................................................... 3

  5.1 Socio-Cultural ....................................................................................................................... 3

  5.2 Technological ....................................................................................................................... 3

  5.3 Economical .......................................................................................................................... 4

  5.4 Environmental ..................................................................................................................... 4

  5.5 Political ................................................................................................................................ 4

  5.6 Legal ................................................................................................................................... 4

  5.7 Ethical .................................................................................................................................. 4

Requirement Specifications ............................................................................................................. 5

  6.1 Functional Requirements ...................................................................................................... 5

  6.2 Non-Functional Requirements ............................................................................................ 5

    6.2.1 Usability ....................................................................................................................... 5

    6.2.2 Portability ..................................................................................................................... 5

    6.2.3 Maintenance & Extensibility ......................................................................................... 5

    6.2.4 Reliability .................................................................................................................... 5

    6.2.5 Security & Privacy ....................................................................................................... 5

Project Methodology ....................................................................................................................... 6

  7.1 Assumptions .......................................................................................................................... 6

  7.2 Research ............................................................................................................................... 6

    7.2.1 Behavioral finance analysis ......................................................................................... 6

    7.2.2 Impact of Mood on Stock Markets .......................................................................... 8

    7.2.3 Literature review ....................................................................................................... 10
List of Figures

Figure 1: Format of tweet upon retrieval
Figure 2: Format of tweets after processing
Figure 3: Format of structure after processing
Figure 4: Plot of Tweets’ length vs Date
Figure 5: Plot of Tweets’ likes vs Date
Figure 6: Plot of Retweets’ vs Date
Figure 7: Table of mood states by POMS and their related words and seed words
Figure 8: Table of Calm mood states with its related words and seed words
Figure 9: Table showing the results of the sentiment analysis
Figure 10: Correlation for the positive mood states using 3 days’ time window
Figure 11: Correlation for the positive mood states using 7 days’ time window
Figure 12: Correlation for the neutral mood states using 3 days’ time window
Figure 13: Correlation for the neutral mood states using 7 days’ time window
Figure 14: Correlation for the negative mood states using 3 days’ time window
Figure 15: Correlation for the negative mood states using 7 days’ time window
Figure 16: Nervous vs Russell 3000 using 7 days tweets
Figure 17: Calm vs Russell 3000 using 7 days tweets
Figure 18: Negative vs Russell 3000 using 7 days tweets
Figure 19: Neutral vs Russell 3000 using 7 days tweets
Figure 20: Positive vs Russell 3000 using 7 days tweets
Figure 21: Plot of prediction vs actual prices given five days tweets
Figure 22: Plot of prediction vs actual prices given seven days tweets
Abstract

The stock market movement is uncertain and dependent on various factors. This makes predicting the stock market one of the most significant applications in business and finance. Numerous forecasting methods have been developed to predict future stock prices. These methods include some of the early attempts such as times series and the statistical analysis, as well as some of the later ones such as the technical analysis or the fundamental analysis, which are two major types of studies used in this field. The technical analysis consists in the analysis of stock prices’ historical data using machine learning. The second analysis is referred to as the fundamental analysis and it is the assessment we will be exploring in this report. This analysis is based on the theory of behavioral economics that explains that emotions can deeply influence decision-making and individual behavior. It suggests that this could be applied to large societies, meaning that they can share general mood states which implies that societies can experience general sentiment phases. By extension, we could think of social media as general public repositories for thoughts and opinions that can be analyzed to predict the stock market. For our project, we collected Tweets which are then analyzed with regard to different mood states to obtain a mood score and a correlation value which are used in our forecasting of the index Russell 3000.
Introduction

Stock Markets are key players in countries’ economies and important operators when it comes to industrial growth; and thus, economic growth. Industry professionals and investors have constant interests in getting insights about stocks’ prices overtime and are regularly financing research regarding the field.

Consequently, stock exchange forecasting, which “is the act of trying to determine the future value of a company stock or other financial instrument traded on an exchange” (Wikipedia) has been in lights of academia and business for many years now. Different attempts have been made to predict its movements, but very few have made significant advancements. Early attempts were based on the Efficient Market Hypothesis and random walk theory. The efficient market hypothesis argues that new information about a company’s prospects is reflected immediately in the according stock price. The random walk theory claims that prices follow a random pattern which makes them unpredictable.

However, there are two issues regarding the efficient market hypothesis. First, studies have shown that stock market prices do not follow a random walk pattern and can in fact be forecasted to a certain extent. Second, recent studies claim that news may not be predictable, but online social media offers early indicators that can be utilized to anticipate changes in different commercial and economic setups. Hypothetically, this might be the case for the stock market as well. Several discoveries have been made trying to prove this hypothesis. For instance, the article “The predictive power of online chatter” proves that online chat activity can be used to predict book revenues, and the article titled “Predicting Movie Sales from Blogger Sentiment” predicts movies sales using evaluations of blog sentiments. Moreover, one cannot deny the fact that news has an impact on prices of the stock market; however, it is also believed that general mood levels are as relevant. Significant progress has been made over the
last couple of years and it is now possible to perform sentiment analysis to access public mood states straight from online social media sources and content.

**Project Scope**

4.1 **Project Overview**
This project aims to provide a stock market prediction using sentiment analysis on posts and interactions such as Tweets which provide a sample representation of general moods and sentiments. Our approach consists in the collection of these Tweets on which we will perform sentiment analysis to extract sentiments reflected by the users to analyze the correlation between the sentiments and the stock prices.

4.2 **Target Audience**
The target audience of this project is the core of professionals and academics researching stock market prediction, especially using sentiment analysis. Students studying machine learning and finance are also invited to take a look at this project as it portrays some of the fundamentals of stock market prediction.

4.3 **Objectives**
By working on this project, we plan to attain several project-oriented and learning objectives. They are as follows:

4.3.1 **Project Objectives**
- Providing a software solution that can perform sentiment analysis on different Twitter posts.
- Developing a program that can model and predict the stock market movement based on deducted general sentiment levels.
- Helping beginners to understand the stock market.
4.3.2 Learning Objectives

- Learning sentiment analysis techniques.
- Understanding the behavior of the stock market and how it can be impacted.
- Understanding the evolution of stock market values and the impact of mood states.
- Learning how to correlate general sentiment states to stock market fluctuations.
- Getting familiar with deep learning frameworks and programming languages used for mathematical processing and machine learning.

STEEPLE Analysis

5.1 Socio-Cultural
The application we are implementing attempts to bring a solution for stock market prediction through the analysis of social media. This beneficial to society in various ways. One of which is the fact that it highlights the importance and usefulness of social media as a source of data for analysis in all kinds. Also, it incites to be knowledgeable about the stock market and emphasizes the need to understand how it works and how it fluctuates.

5.2 Technological
The program makes use of social media posts to come up with a prediction with regard to stock prices. This technology is still in the development process as most solutions cannot provide a reliable prediction rate that can actually be used for considerable investments. Therefore, this solution can be further improved to incorporate other factors and sources to the prediction process.
5.3 Economical
This program offers a solution to make adequate investments based on social media interactions and sentiment analysis since it helps the user assess the stock market movement. Various are the economic impacts of this production. First, since the program can help users make prediction, they would not have to rely on investment companies to make decision from them, and therefore, this can reduce the usual cost of investment one would have to pay in case he or she did not have access to the program. Second, using this program can lead to money earnings, as it would be utilized to invest.

5.4 Environmental
Our project does not have any repercussions on the environment.

5.5 Political
In its nature, stock market prediction is often a politically charged activity. Resources’ allocations, careers and strategic decisions may hinge on the way things are predicted rather than the way they actually turn out.

5.6 Legal
The solution could lead to the creation of new regulations that ensure its appropriate use and that prohibit any malicious party from using it to serve its wrong intentions.

5.7 Ethical
Stock market forecasting raises many ethical issues. In many cases, it is possible to bias a prediction in one way or another by varying the length of the sample which is fitted, by deciding to include or suppress influential observations, by focusing on short-term trends rather than long-term trends or vice versa, by restricting the class of models or the set of explanatory variables which are investigated.
Requirement Specifications

6.1 Functional Requirements
The system shall offer the following functional requirements:

1. The system shall present historical data relevant to the index.
2. The sentiment analysis shall be performed on recent Tweets.
3. The charts shall offer recent data and comparisons.
4. The system is able to perform sentiment analysis on previous data.
5. The system shall predict the stock market movement relative to one of the indexes.

6.2 Non-Functional Requirements

6.2.1 Usability
The system should offer a user-friendly interface. It should include friendly icons and other features which will make a friendly interface for unexperienced users.

6.2.2 Portability
The system should be deployable in other environments and transportable to avoid any losses of dependability in case of a hardware failure.

6.2.3 Maintenance & Extensibility
The system should require minimum maintenance and be extensible.

6.2.4 Reliability
The system should remain functioning in the event of an update.

6.2.5 Security & Privacy
The system shall have robust security.
Project Methodology
We followed the methodology used for academic research projects to write this paper. We suggest that Twitter can be a sample population representing a general public that can be studied based on sentiment analysis to predict the stock market. To prove this hypothesis, first, we conducted a literature review on the topic during which we explored the different strategies used to perform this task. Second, we built our experimental software which attempts to prove our claim. Third, we discuss the results and come up with conclusions.

7.1 Assumptions
• Punctuation is believed not to influence mood scores of users’ posts.
• The number of Tweets collected is sufficient to make a prediction.
• Hyperlinks do not provide any sentiment insights.
• Special characters omission does not influence the prediction.

7.2 Research
7.2.1 Behavioral finance analysis
Behavioral finance specialists have been showing that the stock market’s movement is driven by the psychology of investors since the early 1990s. Investors make mistakes such as other human beings as they are affected by emotion. Observations show that market anomalies exist, which contradicts the hypothesis of the efficient market (Farma 1970) which stipulates that share prices cannot be predicted as every piece of available information is reflected in market prices.

Numerous are the example that portrays these market anomalies. One we could think is the calendar anomalies which consist in the movement of stock market returns based on the season of the year. As a matter of fact, the January effect claims that returns are higher in January, on average, in comparison to other months (e.g., Thaler 1987). This could be explained by tax-loss selling, which is the results of investors selling shares which have had a bad performance during the year to pay less taxes.
Afterwards, in the beginning of the year, the price of shares increases as it bounces back from the selling.

Technicality also intervenes when it comes to market anomalies. The momentum effect asserts that past winners keep on performing well. This effect has been verified studying single stocks (Jegadeesh and Titman 1993) and indices (Chan et al. 2000). According to the efficient market hypothesis, although persistence cannot be predicted, investors have tried to analyze the performance of mutual funds as indicators of future returns (Grinblatt et al. 1995).

Numerous explanations have been provided to illustrate this phenomenon. Researchers have explained both the calendar anomalies and the technical anomalies; however, we will be looking into the anomalies driven by emotions and feelings as they are more relevant to our study.

According to behavioral finance, investors can be categorized into two groups. The former group resolves in the rational arbitrageurs. These investors are well-informed. They do not give up to sentiments and emotions; and their choices and actions are purely backed by logical conclusions. These people are referred to as “smart money” in papers and writings. The latter consists in noise traders. These investors illogically rely on emotion and other irrational information to make decisions regarding buying or selling a stock. This group of people is known to react to news and media.

The efficient market hypothesis suggests a relation between these two groups. It argues that they trade against each other, getting prices back to their original values after many trades. Therefore, we can assume that noise traders influence the market but only for a short period of time. Following is the rational traders’ moves which restore the market’s equilibrium (Fama 1970).

This goes against what behavioral finance suggests. It claims that research has shown that the power of noise traders is greater compared to rational traders’. The positive feed-back strategy, presented by De Long et al. (1990), argues that noise traders can be influenced by other noise traders when selling and
buying stocks. Consequently, noise traders sell in case of falling prices and buy in case of rising prices. This means that the behavior of tomorrow’s noise traders can be anticipated.

All in all, we can say that sentiment can impact the share prices if there is limited arbitrage. Different attempts using sentiment analysis have tried to predict share returns such as trading volume (Baker and Stein 2004), market volatility (Whaley 2000). Different sentiment measures have been proposed in order to forecast share returns, such as investor and consumer surveys (Brown and Cliff 2005; Lemmon and Portniaguina 2006; Qiu and Welch 2004). In this paper, we make use of general sentiment analysis.

### 7.2.2 Impact of Mood on Stock Markets

Neuropsychologists argue that mood is impacted by numerous causes. Serotonin was found to be responsible for negative mood, whereas dopamine is known to mediate the cognitive effects of positive mood (Mitchel and Philipps 2007). It was discovered that that stress levels and events’ occurring are not the only factors that influence one’s mood levels, but social interaction as well.

Many researches have been conducted to showcase examples of mood’s impact on the stock market. According to a study regarding the period between 1927 and 1989, stock returns at New York Stock Exchange are lower on cloudy than on sunny days (Saunders 1993). Hirshleifer and Shumway (2003) came later to prove the effect of weather on stock returns by showing that sunshine correlates positively to stock returns in 26 countries. Both of these studies claim that weather affects mood which influences investment behavior.

Weather is one of the factors that affect mood, but it is not the only one. Sport events are also known to impact individuals’ mood states (Wann et al. 1994). Edmans et al. (2007) showed that international soccer games’ outcomes impact stock returns. They observed that losses of national soccer teams in important competitions lead to a decrease in stock returns. This was shown to be true regarding other sports, depending on their popularity in their given countries. It was also discovered that NFL game
results affect returns of businesses that are headquartered locally (Chang et al. 2012). Sleeping is also known to influence stock returns as well. The daylight-saving anomaly states that Mondays following daylight-saving weekends witness a decrease in stock returns compared to regular Mondays. This is supported by the hypothesis that people shy way from investment as they feel more anxious due to gains or losses of sleep (Kamstra et al. 2000).

These events can be categorized into two groups. Sport outcomes can be considered as short events or single events, whereas the weather effect and the daylight-saving effect can be regarded as continuous outcomes. Both are known to impact individuals’ emotions. These anomalies can be illustrated by the misattribution bias. It explains that individuals take risky decisions depending on their mood (Johnson and Tversky 1983). People experiencing a positive mood state are more likely to be more optimistic when assessing future events. Therefore, individuals’ emotional state and well-being plays an important role in subjective probability estimations (Wright and Bower 1992).

The relationship between the risk tendency and negative and positive mood levels was clarified by the affect infusion model. It argues that people going through positive mood states count on positive hints in order to make choices (Forgas 1995). Also, due to the mood priming phenomenon, risks are associated to positive results when traversing a state of positive mood. In contrast, people traversing a negative mood state would perceive this given risk as loss. We can assume that the risk-taking mindset and decision-making would be more apparent in people experiencing positive levels of mind as they relate risks to earnings. On the other hand, individuals in negative mood states are more reluctant and careful when they take decisions.

The Affect Infusion Model was verified by many laboratory experiments. Yuen and Lee (2003) conducted experiments on subjects by inducing them to positive and negative mood states by showing them respective video clips and measuring their risk-taking tendencies. The findings show that people induced into a depressed temper were less prone to taking risks than those induced into a neutral state, whereas positive moods did not differ in any major way from the neutral ones.
In this section of the paper, we state that mood fluctuations impact risk-taking strategies which in turn affect one’s willingness to take hazardous investments. Therefore, mood levels of market players are expected to impact stock returns’ movement.

7.2.3 Literature review
Prior to the massive increase of online social media data availability that been witnessed during the last couple of years, research relied on exogenous factors as variables of study. These include weather data, historical data, sports outcomes and many more. Online social media have allowed for the analysis of users posts which are indicators of their emotional states. One of the works that made use of this change has been performed by Bollen et al. (2010). They have proven that mood states reflecting the general sentiments behind public tweets could be used to predict the Dow Jones Industrial Average. They were able to prove the correlation of overall happy and calm moods to the Dow Jones Industrial Average. Some other studies combined Twitter data to other sources such as Rao and Srivastava (2012) who included Google search volumes to predict the expected returns, volatility and trading volumes of raw materials and stock prices. Nann et al. (2013) made use of company news and online messages boards as well to create a trading model which can perform better than the S&P index by 0.24%. Other social media, other than Twitter, have been investigated. These include stocktwits.com which is a social media reserved for investors. It was the subject of the study conducted by Oh and Sheng (2011), which touches upon the predictive indications of posts to the stock market’s movement by studying roughly 70 000 micro-blog messages posted over a period of 3 months. Therefore, we can suggest the hypothesis that increased positive social mood states on Twitter correlate to higher stock market returns.

7.3 Approach
It is also relevant to say that the community structure plays an important role when trying to extract sentiment from social media interactions. Evidence shows that information cascades and diffusion
processes are mechanisms that heavily present in the fields of computer science and social network analysis (Granovetter 1973; Hinz and Spann 2008; Kempe et al. 2003; Leskovec et al. 2007).

Experimental study has already demonstrated that mood levels are contagious (Hatfield and Cacioppo 1994). According to Sy et al. (2005) as well as Bono and Ilies (2006), group members and followers are impacted by positive states of minds of their leaders. Also, studies have also made the case that individuals who actively listen to each other transfer emotions from one another (Neumann and Strack 2000). Another instance of diffusion of emotion is the spread of happiness across social media, witnessed by Fowler and Christakis (2008).

A recent exploration was carried on 689,003 Facebook users whose news feed have been manipulated in regard with the volume of positive and negative contents. The experiment indicates that individuals who were exposed to negative contents performed less positive status updates, whereas people who were exposed to positive contents performed less negative status updates. Therefore, one can conclude that Facebook users’ emotions can influence other Facebook users’ emotions. This experiment showcases the relationship between feeds and reactions and verifies that textual contents spread feelings even in the absence of direct social networks’ interactions.

Many researches have investigated Twitter as a network structure. Most of these researches only valorize the huge amount of data can be retrieved from it, but neglect the importance of sentiment analysis on mood and the diffusion of emotions. Twitter users are tightly connected to one another as they retweet each other’s posts, which create a large network of content that can be forwarded and spread.

7.4 Data

7.4.1 Data collection

We collect data from Twitter using Twitter API and using the library “tweepy” of python to help use get data and filtrate them easier.
7.4.2. Data processing

When Tweets are first retrieved, they are loaded as a data structure that is around 9000 to 10 000 characters. It looks like this:

![Figure 1: Format of Tweet upon retrieval](image1.png)

It is a significantly heaving considering the fact that Tweets can only be 140 characters. Tweets are then written to text file which can be used for processing. Afterwards, Tweets’ texts are stored in a Pandas data frame. Once printed, it looks like this:

```
Dominique's story is a reminder to us that when... 
This generation of climate activists is tired... 
And in the U.S., 33-year-old Rukaiyaa Odonho... 
I met 23-year-old #14thHeartbeats in Berlin last... 
They're people like 10-year-old #BlackFloyd... 
Young people all over the world are leading the... 
To all who celebrate today, happy Easter from... 
The attacks on Trump and Easter optimizations... 
Norris Davis is one of the world’s great treasures... 
Congratulations, Tiger! To come back and win t... 
Another good story worth sharing: From one #Ni... 
From a big NBA fan, congrats to future Hall of... 
In just a few minutes, I'm taking the stage at... 
A voice everyone should hear, https://t.co/nu... 
Here’s a story about people doing good that’s... 
Great to see Chicago’s historic mayor ride B... 
Valerie is one of my oldest friends and advisors... 
Last night I had the chance to meet with first... 
Just in the nick of time. My brakes have now... 
And here are some ways to help cyclone victims... 
The floods in the Midwest and in southern Afri... 
In 2013, I visited the city town of Mwengu... 
Michelle and I send our condolences to the peo... 
Ladishindevho and I are rooting hard for Yao... 
What a great moment. Happy Wash, Virginia. h... 
Michelle and I send our deepest sympathies to... 
The Crew Dragons bear on quite a ride since 1... 
Prentis, Alice, and Sefano are just three of m... 
And #Fororodjio is helping the next genera... 
#MoreBible is helping refugees resett... 
```

![Figure 2: Format of Tweets after processing](image2.png)
The tweet ID, the length, the date, the source of the tweet, the likes and the retweets are stored in a numpy data frame. Once printed, here’s what it looks like.

![Table of tweets](image)

### Figure 3: Format of structure after processing

#### 7.4.3. Data Visualization

We also implemented functions to visualize information regarding batches of Tweets:

- **Length vs Date**: Shows how length changes with regard to the time

![Length vs Date](image)

### Figure 4: Plot of Tweets’ length vs Date

- **Likes vs Date**: Shows how likes fluctuates with regard to time

13
Figure 5: Plot of Tweets’ likes vs Date

- Retweets vs Data: Shows how Retweets rate changes given time

Figure 6: Plot of Retweets’ vs Date

7.4.4 Sentiment analysis

7.4.4.1. Sentiment analysis using TextBlob

Prior to the sentiment analysis, textual data is treated to allow for optimal assessment. All special characters and non-alphabetical terms are removed as it is believed that they do not provide emotional insights. Hyperlinks are removed as well as they do not express feelings.

Using TextBlob library, we can score every Tweets ranging from -1 to 1. If the score is positive, the given user’ Tweets is considered to be positive whereas; if the score is negative, the user’s Tweet is viewed as being negative. If the score is 0, the users’ mood is neutral.

Once every Tweets gets a score, we compute the number of Tweets of every given mood state (positive, neutral and negative) and calculate a percentage indicator for every mood, which translates to how many people feel positive, neutral or negative today.
7.4.4.2. Sentiment analysis using the POMS model

To access mood states, we chose to utilize the POMS model. This model has been approved and used in a variety of studies. It was first developed by McNair et al. in 1971 to measure individual’s psychological health. It resolves in six mood states which are: Anger, Depression, Vigor, Fatigue, Tension and Nervousness. This model was later improved by R.Grove et al. who worked on an abbreviated POMS that counts 40 terms and an additional mood scale that is “esteem”. These 40 terms were then extended to capture as many sentiments related to the seven moods as possible. These extension words are referred to as seed words.

<table>
<thead>
<tr>
<th>POMS Mood States</th>
<th>Related Terms</th>
<th>Seed Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td>Annoyed, Bitter, Resentful, Resentful, Peeved, Furious, Grouchy</td>
<td>Outraged, Indignant, Irate, Sullen Mad, Irritable, Enraged, Heated, Bitter, Resentful, Furious, Offended, Annoyed,</td>
</tr>
<tr>
<td>Depression</td>
<td>Discouraged, Helpless, Hopeless, Miserable, Sad, Worthless</td>
<td>Tragic, Discouraging Forlorn, Pessimistic, Pathetic, Desperate, Useless, Helpless, Gloomy, Wretched, Depressed, Sad, Dejected, Despairing, Dismayed,</td>
</tr>
<tr>
<td>Vigor</td>
<td>Full of pep, Lively, Cheerful, Energetic, Vigorous, Active</td>
<td>Potent, Cheery, Lively, Zealous, Animated, Frisky Merry, Rosy, Optimistic, Active, Dynamic, Spirited, Buoyant, Bright, Peppy, Chirpy, Vital Brisk, Perky,</td>
</tr>
<tr>
<td>Fatigue</td>
<td>Bushed, Weary, Worn Out, Exhausted, Fatigued</td>
<td>Drained, Disabled, Weakened, Fatigued, Exhausted, Disgusted, Overworked,</td>
</tr>
<tr>
<td>Mood State</td>
<td>Related Words</td>
<td>Seed Words</td>
</tr>
<tr>
<td>-------------------</td>
<td>----------------------------------------</td>
<td>---------------------------------</td>
</tr>
<tr>
<td>Overtired Sleepy</td>
<td>Bored, Jaded, Bushed</td>
<td></td>
</tr>
<tr>
<td>Tension</td>
<td>Tense, Restless, Anxious, On-edge, Uneasy, Nervous</td>
<td>Apprehensive, Concerned, Antsy, Nervous, Impatient, Insecure, Disturbed, Anxious, Uneasy, Afraid, Strained, Distressed Agitated, Edgy, Tense,</td>
</tr>
<tr>
<td>Esteem</td>
<td>Satisfied, Embarrassed, Proud, Ashamed, Competent</td>
<td>Satisfied, Abashed, Competent, Qualified, Decent Honored, Ashamed, Apologetic, Noble, Content, Embarrassed, Appreciative, Humbled,</td>
</tr>
<tr>
<td>Nervousness</td>
<td>Uncertain, Forgetful, Confused, Bewildered, Concentration</td>
<td>Inattentive Befuddled, Confused, Puzzled, Perturbed, Forgetful, Disorganized, Distracted, Bewildered, Dazed, Addled,</td>
</tr>
</tbody>
</table>

**Figure 7: Table of mood states by POMS and their related words and seed words**

This model was tested and its shows that the nervousness dimension correlates the best to the stock market compared to other mood states.

Also, in the paper “Twitter mood predicts the stock market”, the authors prove that there is correlation between the calm state of mood and the stock market; however, it does not provide a lexicon that can used to track this emotion when performing sentiment analysis on Tweets. So, we will be creating our own lexicon to capture it.

This is what we came up with:
### Figure 8: Table of Calm mood states with its related words and seed words

<table>
<thead>
<tr>
<th>Mood State</th>
<th>Related Words</th>
<th>Seed Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calm</td>
<td>Uneventful, Quiet, Still,</td>
<td>Level-headed, Harmonious,</td>
</tr>
<tr>
<td></td>
<td>Peaceful, Cool, Relaxed</td>
<td>Undisturbed, Placid, Collected,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pleasant, Sedate, Laid-back,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Serene, Pacific, Smooth,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>tranquil, Mild, Slow, Low-key,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Reposing Quiescent,</td>
</tr>
</tbody>
</table>

#### 7.4.4.3. Scoring

The way we will be scoring Tweets given this approach is through the matching of terms in the lexicon to terms in Tweets. When a match is found, it is mapped back to its relative sentiment and its occurrence is increased.

We will be using Person’s correlation coefficient as it is known for its reliability and its wide use in science. It consists in a measurement to test whether there’s correction between variables X and Y. It is defined as:

\[
p_{xy} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 (Y - \bar{Y})^2}}
\]

\[X\ : \ Mood\ score\ depending\ on\ emotion\]
\[Y\ : \ Closing\ price\ of\ Russell3000\ index\]
\[p_{xy}\ : \ Pearson's\ Correlation\ Coefficient\]

Person’s correlation coefficient ranges from -1 to 1:
-1 refers to total negative linear correlation
1 refers to total positive linear correlation
0 refers to no linear correlation

7.4.5 Reasons for choosing Russel 3000
There are various corporations and companies present in stock markets all over the world. Indeed, each economy has different stock indexes. As our study takes on the challenge of predicting the stock market using social networks’ data, our choice needs to take into consideration the available data for the given economy, and by extension for its indexes as well. Considering that Twitter is the platform we will be using to collect data, and that US citizens are widely present on Twitter, we will be analyzing two of the major indexes which are the Russell 3000 index which is an index that tracks the performance of 3000 publicly-owned large companies in the US. This index was also chosen given that most of its organizations are active on Twitter. This suggests that there would be multiple reactions from users’ considering performance, giving place to a rich emotional context.

Experimentation
We tried various date windows in order to evaluate correlation. The goal is to try to make it as close to -1 or 1. If the score is too close to 0, the Tweet is neglected and won’t be utilized in the forecasting.

We used three, four, five and seven days to access the correlation.

Following the sentiment analysis performed on the collected Tweets, we came up with the following results:

<table>
<thead>
<tr>
<th></th>
<th>Three Days</th>
<th>Four Days</th>
<th>Five Days</th>
<th>Seven Days</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nervous</td>
<td>0.734</td>
<td>0.803</td>
<td>0.831</td>
<td>0.668</td>
</tr>
<tr>
<td></td>
<td>Calm</td>
<td>Negative</td>
<td>Neutral</td>
<td>Positive</td>
</tr>
<tr>
<td>--------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td></td>
<td>-0.734</td>
<td>-0.803</td>
<td>-0.831</td>
<td>-0.668</td>
</tr>
<tr>
<td></td>
<td>-0.088</td>
<td>-0.321</td>
<td>-0.345</td>
<td>-0.578</td>
</tr>
<tr>
<td></td>
<td>-0.208</td>
<td>-0.291</td>
<td>-0.511</td>
<td>-0.442</td>
</tr>
<tr>
<td></td>
<td>0.162</td>
<td>0.440</td>
<td>0.546</td>
<td>0.704</td>
</tr>
</tbody>
</table>

Figure 9: Table showing the results of the sentiment analysis

8.1 Correlation Results: Positive vs Russell

8.1.1 Correlation for the positive mood states using 3 days’ time window:

![Graph of correlation between positive mood states and Russell index](image)

\[ C = 0.16199 \]

Figure 10: Correlation for the positive mood states using 3 days’ time window
8.1.2 Correlation for the positive mood states using 7 days’ time window:

\[ C = 0.70356 \]

Figure 11: Correlation for the positive mood states using 7 days’ time window

8.2 Correlation Results: Neutral vs Russell

8.2.1 Correlation for the neutral mood states using 3 days’ time window:

\[ C = -0.20844 \]

Figure 12: Correlation for the neutral mood states using 3 days’ time window
8.2.2 Correlation for the neutral mood states using 7 days’ time window:

\[ C = -0.441967 \]

Figure 13: Correlation for the neutral mood states using 7 days’ time window

8.3 Correlation Results: Negative vs Russell

8.3.1 Correlation for the negative mood states using 3 days’ time window:

\[ C = -0.08752 \]

Figure 14: Correlation for the negative mood states using 3 days’ time window

8.3.2 Correlation for the negative mood states using 7 days’ time window:
Figure 15: Correlation for the negative mood states using 7 days’ time window

From the results, we can say that the correlation of the seven days Tweets is higher if compared to the one of three days’ time window. Consequently, we will make use of the seven days’ time windows for our prediction.

8.4 Correlation using 7 days Tweets: Nervous Vs Russell 3000

Figure 16: Nervous vs Russell 3000 using 7 days Tweets

When calculating Pearson’s correlation number for the nervous mood state, we found that it is close to 0.66811
8.5 Correlation using 7 days Tweets: Calm vs Russell 3000

![Calm vs Russell 3000 using 7 days Tweets](image1)

Figure 17: Calm vs Russell 3000 using 7 days Tweets

When calculating Pearson’s correlation number for the calm mood state, we found that it is close to -0.66811

8.6 Correlation using 7 days Tweets: Negative vs Russell 3000

![Negative vs Russell 3000 using 7 days Tweets](image2)

Figure 18: Negative vs Russell 3000 using 7 days Tweets
When calculating Pearson’s correlation number for the negative mood state, we found that it is close to -0.66811

8.7 Correlation using 7 days Tweets: Neutral vs Russell 3000

![Figure 19: Neutral vs Russell 3000 using 7 days Tweets](image)

When calculating Pearson’s correlation number for the neutral mood state, we found that it is close to -0.44196
8.8 Correlation using 7 days Tweets: Positive vs Russell 3000

![Figure 20: Positive vs Russell 3000 using 7 days Tweets](image)

When calculating Pearson’s correlation number for the positive mood state, we found that it is close to 0.70356.

Results and Discussion

9.1 Prediction approach

To forecast the stock market, we will be using the following formula:

\[
f(x) = \begin{cases} 
1, & \text{if } \sum_{i=1}^{N} X_i \times \text{correlation coefficient of } X_i > 0 \\
-1, & \text{else}
\end{cases}
\]

\[x_i : \text{score of mood}_i\]

if \(f(x) = 1\) \(\Rightarrow\) The stock market is supposed to go up

if \(f(x) = -1\) \(\Rightarrow\) The stock market is supposed to go down
9.2 Prediction using five days Tweets vs actual prices

Figure 21: Plot of prediction vs actual prices given five days Tweets

9.3 Prediction using seven days Tweets vs actual prices

Figure 22: Plot of prediction vs actual prices given seven days Tweets
We can say that using seven days Tweets, we were able to achieve better accuracy then when using five days Tweets.

**Technological Enablers**

- **Python:**
  
  Python is the high-programming language we will be using for this project. It allows for a variety of object-oriented development. It is optimized for statistical analysis, sentiment analysis, stock exchange analysis and financials application due to its versatility. It provides various library that provide high-level functionalities.

- **Twitter developer platform:**
  
  The Twitter developer platform is a portal that allows for server-user communication through an interface. It is used for API access and Twitter applications management. To use this portal, an application is required. For this project, we had to submit a request for access where we were asked to details the project, we will be using the platform for. Once the application was approved, we were provided with Twitter credentials which consist in an access token, an access token secret, a consumer key and a consumer secret. All of these login credentials are needed for logging in to the API.

- **Tweepy**
Tweepy is a python library for the Twitter API. It allows python to communicate with the Twitter platform. It also supports Twitter access using authentication classes and provides a stream listener that can retrieve Tweets. This enables us to use Tweets for our project.

⇒ Numpy

Numpy is a python library used for computation. It offers high-performance multidimensional array objects that can be used to perform heavy computations. It also provides tools for using these arrays and working with them. In our project, we use its data structure to store Tweets and so that they can be used in our computations.

⇒ Beautiful soup

Beautiful soup is a library developed for python that allows for the retrieval of data from XML and HTML files. It was used in this project to access data relative to the Russell 3000 index.

⇒ Pandas

Pandas is a software library written for the python programming language. It provides functionalities for data manipulation and analysis. One of the many functionalities we used during our implementation is the operations offered by Pandas to manipulate numerical tables.
Textblob is a python library. It was developed to enable users to process textual data. It offers a simple API for natural language processing, noun phrase extraction, sentiment analysis, classification, translation, part-of-speech tagging and more.

Requests is a python library that was developed to allow users to interact with the Apache2 Licensed HTTP library. It provides high-level modules that simply access to through automated additions of query strings to URLs and the avoidance of form encoding data.

Matplotlib is a python library that offers a collection of functions that are similar to those offered by mathlab. It is mainly used for plotting 2D graphs. This makes things easy for plotting by allowing formatting axes and providing font properties and features to control line styles.
**Feasibility Study**

Predicting the stock exchange is a very complex and challenging task. There are simply too many future variables and events that come into play to account for. The stock market is the field of play of sellers and buyers, and its movement depends on offer and demand. More buyers and the price goes up, more sellers and the price goes down. Transactions are results of decision-making processes that are influenced by emotions, feelings and intuition. These are factors that are very difficult to quantify and model; and therefore, to predict, which makes the stock market hard to predict as well. That is why this project required an analysis of the correlation between public moods and stock returns. Extensive research was also required to understand how the stock exchange is influenced by traders’ decision-making processes. Also, since Twitter is a huge platform, and the Twitter Api also allows for a limited number of the Tweets, we can only access the general sentiment of a sample population, that is generalized to the whole platform. Our prediction is viable given all the parameters that it accounts for, but the stock market is a very hard entity to forecast.

**Conclusions and Future perspectives**

Due to time constraints, we were not able to finish the development of the project to the point we were hoping for. In case we are given the chance to work on this project again, we would like to incorporate various other features and functionalities, as well as considered different indicators that might be relevant to our prediction.

Although our solution is viable, it does not attain a high accuracy. A higher accuracy would yield better profit; and it is thus one very important objective to take into consideration. Also, we would like our program to perform a better sentiment analysis by including a lot more tweets. This would require a premium access to the Twitter Api and allow for the extraction
of a more generalized general sentiment. Moreover, one of our main targets is to be able to predict financial indexes and markets other than the ones in the US. Even though the United States’ financial market is a good market to study in the way that it provides a lot of companies to work with and that sentiment related to it is widely available on the web, it is very competitive in the way that numerous predictions are performed on it, which makes chances of making profit very small. Other markets in Europe and Asia also provide many companies, and enable traders to make better profits as less predictions are done on them. Therefore, one of our many goals would be to perform sentiment analysis on tweets in other languages than English.

During Summer 2019, we will be working on this project to enhance it. This work will be done in collaboration with CDG Capital to which this project was introduced. The vision is to capture emotional responses related the Moroccan Stock Market. The hardship is that very limited interactions and posts can be analyzed to make a prediction. The challenge is to introduce other sources of social media interactions such as Facebook and news article in order to capture sentiment.
References


S. Asur and B. A. Huberman 2010 Predicting the Future with Social Media arXiv:1003.5699v1


