# Social Mood Impact on Financial Decision Making: A Study of Twitter Sentiment on Stock Index Volume

Submitted by

Gustavo Porras

**Master of Science in Finance**

**National College of Ireland** **2019**

**Abstract:**

Emotions and human behavior are an important subject for academics in the financial field as emotion and mood haven proved by psychologist as a main factor in people behavior and decision-making process. Hence, the modern improvement in social media like Twitter has open new possibilities for exploring more the emotional polarity in human, and their perception toward companies. Does this emotional public sentiment affect people's investment decisions? In this dissertation we explore whether the Nasdaq Composite Index (IXIC) volume is affected by the public mood, using Twitter sentiment analysis and how we can use that information for modelling Nasdaq Composite Volume Change. Our finding prove that sentiment is correlated with change in the IXIC volume, but we can use that information predict change in the IXIC using time series with small dataset.

# Submission of Thesis and Dissertation

**National College of Ireland**

**Research Students Declaration Form**

***(Thesis/Author Declaration Form)***

**Name: _____**

**Student Number: _____**

**Degree for which thesis is submitted: _____**

**Material submitted for award**

(a) I declare that the work has been composed by myself.

(b) I declare that all verbatim extracts contained in the thesis have been distinguished by quotation marks and the sources of information specifically acknowledged.

(c) My thesis will be included in electronic format in the College Institutional Repository TRAP (thesis reports and projects)

(d) ***Either*** *I declare that no material contained in the thesis has been used in any other submission for an academic award.

***Or*** *I declare that the following material contained in the thesis formed part of a submission for the award of

_____

*(State the award and the awarding body and list the material below)*

**Signature of research student: _____**

**Date: _____**

# Acknowledgements

Firstly, I would like to acknowledge and thank my girlfriend for all her support and love that she has given me over the past five years. I would like to thank my supervisor, James Mccloskey for all his advice over the dissertation process. I would like to thank specially to my friend, Daniel Mellado for all the help and his invaluable support and unconditional friendship.

# Table of Contents

# List of Tables

# List of Figures

# Chapter 1

## 1. Introduction

### 1.1. Stocks, Indexes, and Information

Predicting stock market moves has captured the imagination and determination of investors, academics, and the general public. Stock market traders are eager to understand the main factors that may influence the volume and impact on the stock's price per day, by, essentially, understanding how the security market might behave, and how individual assets may relate to one other. When they trade financial instruments, in the US, like derivatives, in an exchange, such as the S&P500 or the Nasdaq100, it would give traders a major advantage to potentially maximize their profit. The volume of trades, or the depth of the market, plays a central role in the information channel and the relationship with the stock's price, (Karpoff, 2006).

There are two principle theories about stock price formation. Firstly, the theory of Fundamental or Intrinsic Value Analysis (IVA). Its main assumption is that an individual stock past price behavior will tend to appear again in its future price movement. This is significant as it implies there should be a way to predict future price movements by looking into the past return behavior. Notwithstanding that, Eugene, (1965) found consistency in the thesis that successive changes in the stock prices, between periods, are independent.

Secondly, the Efficient Market Hypothesis (EMH). This theory states that the market is efficient according to an information set, wherein if a certain price "fully reflects" that information set. All available information is already priced into the stock. The EMH assumes independence in the price movement, a crucial difference. Market participants would not be able to perform inferential conclusions about the future asset price by looking only at past price performance, as the price action follows a random pattern, Fama (1970). However, there was a small focus of attention on the trading volume that supports market exchange classic theory. Recent studies describe that, if we do study the volume of trades per day, we might be able to infer a little more information about the asset price, because a positive coexisting relationship between volume and price can be found (Tiep & Mehmed, 2009).

Recent studies highlight the role of emotions in an investor's decision-making process but also emphasized that the general level of optimism or pessimism is correlated with the trading volume and the stock market trading levels, Nofsinger (2005). In relation to the stock market, improvements in machine learning and the Application Programming Interface (API) technology help explore what Nofsinger (2005) called "social mood". Bollen, Ma, & Zeng (2010), used an opinion finder to identify the emotion polarity in Twitter, either positive or negative, to predict a change in the Dow Jones Industrial Average (DJIA). They found a positive causality link in Twitter feeds, between the variables of Dow Jones trading volume and the social mood toward the stock market industry. However, they could not conclude that the social mood alone affects the Dow using Twitter feed historical data.

In this paper, our analysis will be twofold. Firstly, we will perform linear regression and Pearson correlation tests for analyzing the link between volume on the Nasdaq Composite Index and Twitter polarity obtained from psychsignal.com. Secondly, we will mine daily tweets from the twitter's platform using the Python's twitter API library "tweepy", and with that do real-time measures of sentiment with the simple Python API library for extracting common natural language processing (Textblob). After which, we will apply the person's correlation test to measure whether the social mood could predict movement in the IXIC volume.

### 1.2. Dissertation Direction

This specific study will aim to explore whether the social mood affects the traded volume on the Nasdaq Composite Index (IXIC) by analysing Twitter feed activity and the polarity of them toward the IXIC. The research's target is to contribute to the increasing literature around the Social mood and its applicability in the stock market, through the analysis of social media sentiment. As of yet, there is no clear path to follow, as social media sentiment analysis from a financial perspective is only now currently under study. However, this dissertation covers one question followed by two hypotheses. The dissertation question is:

**"Can Twitter feed analysis assist and ultimately predict an existing correlation between online social mood changes and the daily traded volume on the Nasdaq?**

This is based on the theory that new information, along with emotions, plays a fundamental role on a trader's decision-making process. This would lead to pricing anomalies in the stock-market not related to the stock's fundamental value, (Bollen, Mao, & Zeng, 2010).

When we look into an aggregate emotional tendency, about the sentiment on twitter, we expect to be able to infer a general mood or opinion about a particular subject like the Nasdaq Index. According to studies, an increase in the social mood can influence professional investment decision making quickly, (Anchorage, 2014), (See-To & Yang, 2017) and (Agrawal *et al.*, 2018). This can lead to changes in the underlying stock price, far from its long-term fundamental value. This so-called "social mood" affects investors rational behavior. If the market participant shows a good or positive mood, they tend to be more optimistic toward certain events. This would increase the likelihood of investing in risky assets, like stocks. The contrary tends to happen when the mood is negative towards a security. We will explore this effect based on the assumption that each tweet represents an individual opinion. Therefor an aggregate example should provide an accurate representation of the public social mood (Bollen, Mao and Zeng, 2010). *Table 1. 1* showing the two hypotheses are outlined below:

*Table 1. 1 Hypotheses and Sub-Hypotheses*

| **Historical Sentiment Analysis using Twitter Feed Hypothesis** | |
|---|---|
| $Ha1$ **=Hypothesis 1** | There is a significant correlation between Twitter Sentiment toward the Nasdaq Composite Index and the Nasdaq Index Volume for a given day. |
| $H01$**= Null Hypothesis 1:** | There is no significant correlation between Twitter Sentiment toward the Nasdaq Composite Index and the Nasdaq Index Volume for a given day. |
| **Short Window Analysis using Twitter Feeds** | |
| $Ha2$ **=Hypothesis 2** | By modeling short daily Twitter sentiment volume and stock volume for the Nasdaq, next day prediction on Nasdaq direction can be made. |
| $H02$ **=Null Hypothesis 2** | By modeling short daily Twitter sentiment volume and stock volume for the Nasdaq, next day prediction on Nasdaq direction can not be made. |

Regarding the first hypothesis, we intend to explore the link between the Nasdaq composite index volume and the Twitter Polarity sentiment (positive/negative), with data obtained from the psychsignal.com platform. We will use linear regression, correlation and an analysis of variance ANOVA to test our hypothesis. Regarding the second hypothesis, we will use daily Twitter activity toward the Nasdaq index and its volume to make an inferential prediction about the next day Nasdaq index's movement. We will extract Twitter's feeds from the Twitter platform and then analyze those Twitter's feeds, using the Python's library twitter API access "Tweepy," and also the "Textblob", a simple API library for extracting common natural language processing (NLP).

# Chapter 2

## 2. Literature Review

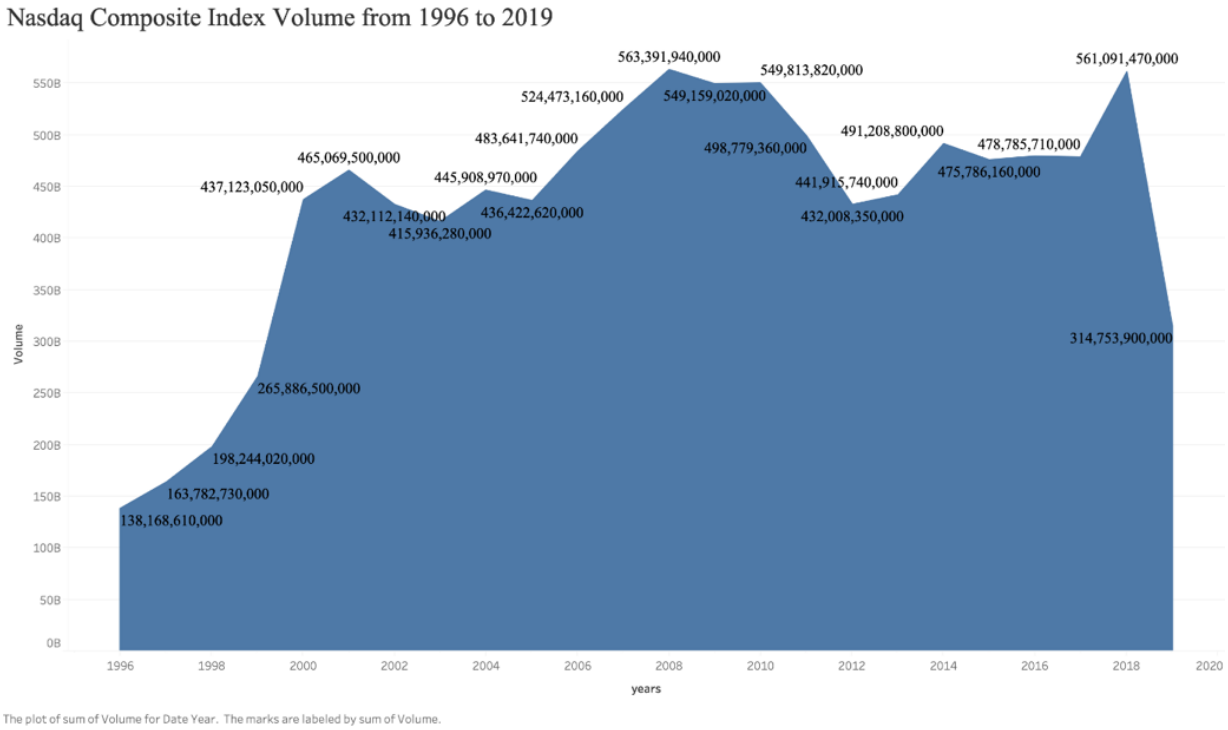
### 2.1.Background and Scope

This chapter details the theoretical background to support our research question. Traditional theories that highlight stock price behaviour, such as, the Fundamental Value Approach (FVA), and the Efficient Market Hypothesis (EMH), will be considered. Additionally, the criticisms of both models will be explored. The 'social measure' as a market sentiment tool will be introduced and explored. Focusing on the social impact of Twitter, its scope and penetration levels will be investigated to make a strong case for its potential impact on the financial landscape and index volume.

### 2.2. Financial Markets

The financial market is a global market where companies, financial institutions, governments, and individuals converge in order to provide or acquire fresh liquidity to support their growth targets in the short and long term. This capital is obtained from investors who also seek to increase their asset wealth by investing in a large number of financial instruments. Financial instruments, broadly speaking, are intangibles assets, stock, bonds, derivates, indexes. These are expected to deliver some positive return or profit in the form of tangible cash-based positive returns (Darskuviene, 2010). The global financial landscape is a rich and complex environment, but for our purposes, we will focus on the price behavior of the technology-heavy Nasdaq composite, based in the US. Millions of traders, investors, and market participants converge every day, trading billions of stocks in a massive assets market. *Figure 2. 1* below shows the Nasdaq Composite Index trading volume, 1996-2019.

*Figure 2. 1 Nasdaq Composite Index Volume, 1996-2019 (Source: Yahoo Finance, 2019)*




The Nasdaq Composite Index of 2019 is a weighted-index gathering over 3300 common shares exclusively listed on the Nasdaq stock exchange. The index contains international and US companies (Nasdaq, 2017). It is highly weighted on the technology sector with companies, like major multinationals Microsoft, Amazon, Apple, and Facebook. These four companies represent about 28% of the total index's value (Nasdaq, 2019). The Nasdaq Composite Index is a powerful indicator in the financial market, and most market investors follow its price and volume behavior closely. We will focus on the volume changes, as the numbers of shares traded has been modeled in some previous empirical research to explore traders' behavior (Shiller, 1999). Figure 2.2.1 above shows the Nasdaq Composite Index capital magnitude that the index moves every day. Trading volume on the Nasdaq Composite Index (IXIC) represents a valuable piece of information as millions of transactions are transacted. The volume in 1996 was 138.2 billion transactions or approx 3.8 billion shares per day and this value growth significantly the past 20 years. The number of transactions traded in 2018 was approx 561 billion. The index per se is a strong indicator a billions of people convengen in it every day, with the appropriate tools we could make inferential analysis about the market especially regarding the trader's behavior.

In the next section, we will review some of the researches that have used the Nasdaq Composite Index and the historical information, such as volume for empirical financial analysis.

## 2.3. Market Volume

A market volume refers to the total number of shares traded, divided by the total number of shares outstanding, Wang et al. (1993). Trading volume turns out to be an important aspect of interactions within the stock market. However, the asset markets' literature has focused primarily on the asset's price formation. Little work on trading volume has been done, Daves (2003). With the Nasdaq similarity, there have been many studies trying to understand the indexes-price-formation behaviours, but a relatively little analysis on its trading volume. One of the main analytical studies involving volume is related to its ability to measure the reactions of traders about certain informational events, Picus (1983). Volume data can help to understand why some market participants tend to wait until important information is revealed, to make any important decisions. This phenomenon was called "the disjunction effect" by Tversky and Shafir (1992).

Trade volume has been used for exploring some behavioural phenomenon in the stock market, which goes against the rational behavioural approach, such as the Regret Theory. According to this theory, some investors hold stocks that have gone down in value and rush to sell securities that have appreciated in value. Traders rush to sell the stock that has gone up in value because they don't want to regret failing to do so if the securities falls in value, Shiller, (1999). The trading volume can be seen as a key raw input when it comes to exploring behavioural studies in the stock market.

Studies suggest that the volume contains important information about the asset returns and could be used to predict return volatility, Jiranyakul (2007). Tiep & Mehmed (2009) suggest that the volume might contain an informational element that is advantageous for the market's participants in modelling volatility, and traders should include volume in their volatility modelling process. This "informational element" is a broad aspect that must be defined properly for further academic exploration. In this regard, we aim to add understanding about the Nasdaq volume behavior by exploring its link with the public sentiment perception or polarity. Twitter sentiment analysis will be our main tool.

## 2.4. Fundamental Value

Many academics and professionals have undertaken empirical studies to develop techniques to predict movements in stock's price as a predictor of its possible future price behavior. We can identify two fundamental theories regarding a security's price formation: the chartist and fundamental analysis theory and the random walk theory as supported by the Efficient Market Hypothesis (EMH). The chartist and fundamental analyst rely on statistical tools and financial graphs such as moving averages, Fibonachi sequences, bars, lines and the Japanese version of the stock's chart the "candlestick". These contribute to the empirical analysis of market behaviour, and also evaluate if the individual stock price is overvalued or undervalued, Garcia (2017). In other words, they attempt to identify whether the actual stock's price is beyond or under its "intrinsic value". This is the underlying value based on fundamentals without irrational effects, Kim (no date). What is more interesting is the price assumption that most

of the chartist and fundamental theories are based on. Fama, (1965) states that all the chartist techniques make the same assumption. Past behavior of the price and the future price are connected. The historical price contains a valuable amount of information hence investors could infer about the future stock's price behaviour, due to certain "patterns" repeat themselves.

The notion that a stocks historical price is connected, turns out to be the foundation stone of the fundamental analysis, as they state investors can identify patterns through historical stock data price for current investment decision-making.

Parracho, Neves, and Horta (2010), combined traditional chart pattern and trend recognition techniques with generic algorithms. They concluded that once a stock's price or volume behavior is identified, we can combine those results with innovative techniques, using algorithms to support the decision to buy or maintain an underlying asset.

Velay and Daniel (2018), used a similar approach for analysing candlestick charts. They identified a single candlestick pattern like the bearish flag and then applied a deep learning-based recognizer (long short-term memory), and hard-coded algorithms (conventional neural networking) to the pattern found. Despite that, their findings weren't empirically and statistically strong, as in the end they added a new perspective to analyse stock's price formation.

Fundamental analysis along with the chartist theories, start from the idea that the stock could be far from its intrinsic value. By understanding this very logic, we can gain good insight in to the market and its upcoming possible price patterns.

### 2.5. Efficient Market Hypothesis (EMH)

In contrast, according to Fama (1965), the idea that the future price of any security will follow a random pattern. We would not be able to use historical price behavior to infer the future price level of security because the past price and future price are independent. This is known as "random walk" theory and is one of the hypotheses used by Fama to developed his "Efficient market hypothesis (EMH)."

The Efficient Market Theory proposes that at any moment in time, the underlying security price reflects all the available information. He suggests that stock prices were the result of randomly generated noise, and he refers to this "noise" as psychological factors unrelated to political and economic events.

The theory has a traditional economic market approach, where there is an unobservable force that helps the demand and supply of an asset reach the market's price equilibrium around the

securities intrinsic value. It proposes that current prices are independent of previous prices, and rational individuals can buy and sell stocks freely. In other words, the markets behave rational and efficiently due to the force that converges in it.

Fama utilised two main assumptions from the random walk theory. The first is that successive price change are independent of any past changes, and the second is that price changes follow a probability distribution which is not linked to previous price distribution. The two conditions would make sense if the degree of dependence between the past stock prices and the present prices are so low that the investors will not be able to make an accurate prediction of the future security price path. The model is an adequate description of reality because, under the scenario described above, individual investors would not base any investment decision using a history of price moves, for making returns greater than any buy and hold model. In other words, the probability distribution of the price change during period $t$ must be independent of the sequence of price changes during previous periods $t-1$.

The most important assumption in Fama's model is that a market is said to be "efficient". The security price fully represents all information known from the past, the present, and future, potential at any given moment in time. This 'efficient market' is the basis of the Fama model and "fully represents" the current asset price. But is the market really efficient?

Notwithstanding, the recent data available, statistical techniques, learning machines, and data analysis methods, empirical investigations have shown that there are more disagreements about the "*market efficient*" concepts. Sewell (2012) states that the academic community argue that the EMH is unbalanced and impossible to replicate in the real financial space. For the market to be truly efficient, and for prices to "fully represent" information, the model must determine investors risk preferences as well. The model does not define these properly, and it also gives some room to refute its hypotheses empirically. To this end, Beja, (1977), concluded in his "limits of information of market process" research, that the fact that the prices transfer information to investors, is not enough to set the presence of a hypothetical equilibrium price-function. Price, he argues, is a complex resolution process in, where other information could alter the trader's resolution process so the informational condition from EMH was not enough to validate an equilibrium model. Berja was highlighting the significant problem caused by asymmetric information in investors financial decision making.

Grossman and Stiglitz, (1980) argued that the market is not efficient because the harvest and understanding of relevant information is costly. Information quality varies from individuals who could maybe afford more high-quality information resources than those who can't, and disseminating correctly may make for better-informed decisions for the traders less informed, or with lesser resources. Therefore, the price cannot reflect all the information available.

Tirole, (1982) arrived at similar conclusions. Although he highlighted the fact that information might concern other traders' behavior. The stock price would not reflect the informational effect as soon as information arises.

Roberts, (2003) suggests that the collective judgment of investors would make mistakes as long as the security market exists. Pricing anomalies and predictable patterns could appear and could be identified and even persist for a short period of time.

The EMH deserves its place in academic literature. Once applied to the real world of financial marketplace, several of its key assumptions are found wanting and need adjustments for social impact, and also, individual financial behaviour that seems to fundamentally impact trader sentiment and therefore asset price patterns.

### 2.6. Sentiment Analysis

Social networks can enable academics to accurately measure mood and sentiment fluctuations by analysing people's statement and comment on their emotionss or well-being. Rechenthin, (2015) defined sentiment analysis as a technique to understand people's opinion towards a certain topic or theme. It is a process of analysing a piece of online text, or content, to identify the emotional perspective or attitude towards a great variety of topics, such as brand, individual securities, economics events, etc, by using natural language processing (NLP).

Rechenthin, (2015) emphasized that we would be able to classify people's posts as negative or positive, using artificial intelligence. Public sentiment is the public's opinion towards something in particular. For example, the public's opinion toward a certain company or a certain brand. Sentiment analysis is the only tool that we use for mining and processing that perception, in order to discover the practical utility of the public's opinions, Anwar Hridoy et al. (2015). Thus, one person's opinions wouldn't make so much noise, but if we collect and process a reasonable quantity of them, we will get an aggregated indicator called "social mood."

### 2.7. Social Mood

Social Mood as a conceptual term, adapted to the behavioural theory and financial field, tends to describe the general public's level of optimism or pessimism towards a topic. Ziembinski, (2015) argued that the social mood is an aggregated emotional state derived from complex human interactions. Ziembinski, (2015) also claimed that social mood affects the individual persons perception.

Anchorage (2014) described how social mood had been the main subject of study by behavioural finance academics, for exploring trader's behavior and decision-making process in the financial market, especially the security exchange market. He describes that when the general social mood is high, it would drive people to make a hasty investment decision, hence the stock market would be inflated.

We can find one explanation of this phenomenon, in the book *"Descartes's error emotion, reason, and the human brain"* written by Damasio, (1995). He emphasized that the notion of mood was built on emotional people's discernment of their surroundings. Thus people's moods trigger a mixed collection of emotions, such as happiness, sadness, indifference, and fairness, known as "background feeling." This often drives people to inappropriate places in a decision-making situation. Damasio, (1995) provided some insight into how we would measure or evaluate people's moods, from a clinical perspective with reference to some chemicals that can alter the mood and the emotions.

Current improvements in technology, in the learning machine, natural language processing, and application programming interface (API), along with social media, opens new possibilities to link human nature with finance.

### 2.8. Extracting Social Mood

To understand the different techniques used to extract the social mood, we must define the different communicational and programming software that allows academics and institutions to mine the data, necessary to study the social mood or general public sentiment.

Firstly, we have the Application Programming Interface (API). The API is a code that allows communication among different software or application components. The API communication protocols allow developers' application to communicate with other servers (Perry, 2017). Twitter has its own APIs, which allows developers to search tweets, filter real time tweets, account activity, direct message, and also explore Ads activity (Twitter, 2019). There are some metric restrictions to use the Twitter API, but we will come back to this aspect in the next chapter. The most important feature of this type of application is its ability to interact with machine learning for natural language processing.

In the See-To and Yang, (2017) paper, they used machine learning to study what they called "the Market Sentiment Dispersion," and offer proof of its effect on the stock market volatility. They used a regression model of Support Vector Machine (SVM) and managed to explain some change in the stock volatility, using sentiment analysis. Unfortunately, their conclusion about the effect understudy upon the daily changes in return was not conclusive.

Nofer and Hinz, (2015) used a different approach than See-to & Yang, as they quantified the social mood and the contagion among followers to prove that the increase in the social mood levels leads to increase in the stock return. In other words, they used this approach to prove that social mood and stock return are positively correlated. They also used data from the Twitter API, which enabled them to extract the Tweet ID along with other features like time publication, information on followers, re-tweets, and the text content. Their approach was to use this information to construct a Social Mood Index (SMI), which is the sum of positive and

negative tweets. They found some evidence that weighted social mood level can predict stock returns. An increase of 1% of social indicators would lead to 3.3 basis points increase in the German DAX performance index, in the next day, during the period under study. They suggest that one possible explanation of this finding might be emotional contagion among users.

Ranco *et al.*, (2015) also studied the effects of Twitter sentiment on the stock returns. They used a different experimental set-up, statistical models, like Pearson Correlation and Granger Causality to prove the forecasting influence of the Twitter series. Their methodology was based on connecting the movement of companies' stock, with specific events and thus collecting data around those events. So they could make statistical validations to test their hypothesis. They found some interesting dependence between the price of the stocks and the sentiment on twitter toward those stocks.

Azar and Lo, (2016) also studied the social media reaction toward specific events just like Ranco *et al.*, (2015), Nofer and Hinz, (2015) but they used the Federal Open Market Committee (FOMC) meeting as a trigger event. They collected data using the Topsy API which mines English language tweets, mentioning Federal Reverse Governors "Bernanke" or "Yellen" names during the period between 2007 and 2014. Additionally, they also used the python web natural language toolkit Called "pattern" to scrap data. The tool's statistical approach helped them to identified expression. We wills use a similar python's toolkit called "Textblob"; this tool would allow us to process textual data from Twitter tweets, like part-speech tagging, noun phrase extraction, classification, and the polarity or sentiment content in the 140 tweets character (Loria, 2018).

Some studies have focused on isolating certain words to develop a slightly more standardized method. Pak and Paroubek, (2016) studied the number of times that special characters appear in tweets such as: ":-)", ":)", "=)", ":D" etc. for building frequencies. Those sorts of characters generally express happiness or positive emotions. Additionally, they used a classifier for data extraction built on the Naive Bayes classifier. Their principal objective was to build a toolkit to only extract data in English. They tried to offer an alternative method for extracting and classifying raw data. We will clear special character from our tweets using an algorithms script in the python code, for data extraction. The historical data obtained from psychsignal.com has not got any special character to consider. Thus, we programmed our code to avoid taking into account those sorts of characters, as we are interested in the polarity of the word.

There is currently no learning machine method that investigated this phenomenon in different languages. It would be very interesting to have a method with these properties, since, as we know from other studies our background and our race influences our emotions, and the way we make decisions. We could find more information about this topic in the Kumar, Page, and Spalt, (2011), Kumar, (2009) and Kumar, Page and Spalt, (2016) documents.

Rout *et al.*, (2018) refer to the use of the Multinomial Naive Bayes (MNB) to analyse unstructured social media text, but they also use Support Vector machines to identify the sentiment of tweets. By combing two different machine learning algorithms, they were able to

analyse 60195 tweets. They labelled 11,001 as negative tweets and 9485 as positive tweets. The rest of the tweet were classified as neutral, which are tweets that have both positive and negative words or even factual words. While there are a lot of techniques to mine and extract sentiments from tweets based on their lexicon, there is no reliable tool to analyse complex text that has both positive and negative words or factual language. However, the authors suggest that you can create a system to analyse terms and specific sentiments based on a support vector Machine, because this algorithm works better with unstructured media text.

Because social media represent various forms of consumer-generated content (CGC), it has been used to study consumer behavior and business. Xiang and Gretzel, (2010) suggest that networking sites are becoming the primary online source of travel information, since it has improved the possibilities of travellers to analyse past experiences, broadly comparing resources, thanks to the change of information search settings.

Broadstock and Zhang, (2019) analysed an alternative method to study the relationship between social media-based sentiment intraday factor and stock price. Despite the fact that their results were quite mixed, and they did not get decisive results between the variables, the expansion of the study emerged using a broader database and using non-linear regression.

Laurell and Sandström, (2017) and Chen *et al.*, (2014), used similar methods to study the effect of the sharing economy and the value of the stock opinion transmitted through social media. However, Chen and his colleague managed to get a bit more solid results, since they could determine a strong relationship between aggregate search frequency of securities tickets in google and trading investors. Laurell and her colleague stressed some issues that are still to be resolved between taxation and regulation, but using collected user-generated content the data ‘’noise’’ didn’t help them to find more solid results.

Internet, along with mobile phone technologies, are the main platforms for social media, proving creating a perfect environment for human interaction through a considerable amount of social applications, such as Twitter, Facebook, Instagram, etc. The interactive communication generates a stunning amount of raw data that companies and academics are using in order to further explore the human behavior in finance, Zeng et al. (2010).

### 2.9. Social Media and Metadata

The public sentiment and opinion about the security exchange are different in those traders who handle above-average information, than among those who are less informed. This raises the question about the role of public sentiment and social mood in the market, where one should be highly efficient and rational and not subject to emotion or irrational behavior.

The question is no longer whether emotions affect stock market valuation, but how to measure and quantify that effect by combing alternative data from social media databases, like Twitter, with stock-market price feeds, See-To & Yang, (2017).

As Venezuelans, we have witnessed the potential, evolution and impact of what we know today as social networks in 2007. When I was studying at the University of Carabobo, there was a huge student revolution on the streets and on social networks against the 2007 Constitutional reform, that the Government was proposing to make the country a full Communist state. Microblogging websites, like Twitter, allowed people against the reform to spread their ideas and to share content, videos, and small pieces of texts.

The struggle took another perspective, an online social media perspective. Twitter brought people together as a community. Content, comments, and stories shared on the Twitter platform produced a huge International and National impact in favour of the rights and desires of its citizens. Social media offers an unrestricted, unlimited, and more efficient way to share information through a wide range of media. Today global sites like Facebook, Twitter, LinkedIn, Instagram and Reddit all have their own footprint and offer their own platforms. The virtual environment in which they operate captures up to the second changes and moods in modern society and impacts how we all behave, think, react, earn and do business, Yazdanifard *et al.* (2011).

Datafication is one of the most powerful features of social media since people are more intertwined to the informational and content generation process, as reported by Zeng *et al.*, (2010). The ability to make a data-driven decision is fundamental for any organization or business willing to increase their profit adapted to the current modern challenges. However, Zeng *et al.*, (2010), suggest research, based on web-based applications, require a more efficient analytical framework to extract all the useful information or data generated through the virtual means.

The exponential explosion of the internet all over the globe and the improvement of its accessibility has contributed to the positioning of social networks as the main communication system between individuals, from all parts of the world. *Figure 2. 2* shows the internet usage global distribution by regions until June 2019.

*Figure 2. 2  Internet Usage by Region, 1996-2019 (Source: (Internet world Stats, 2019)*

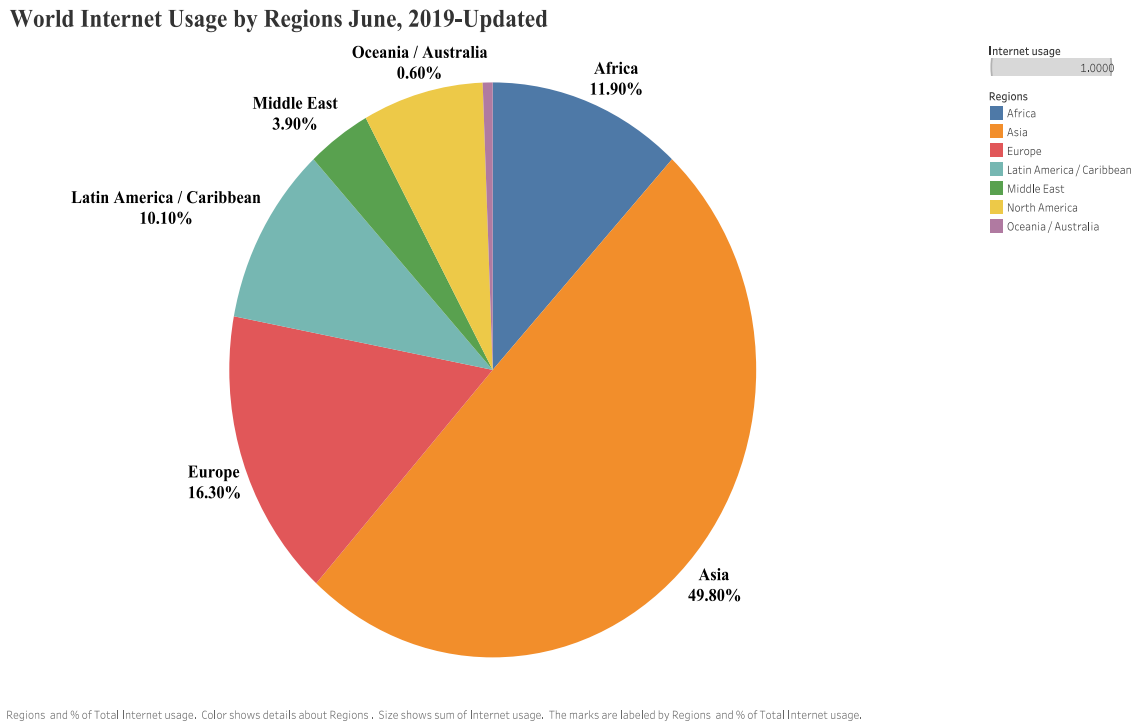



*Figure 2. 2* is the pie chart detailing the global internet usage distribution.  Asia and Europe are the regions that provide the most global internet usage.  Africa has 11.90% of the internet usage following by Latin America/Caribbean with 10.10%.

Social media outlets revolutionized communication and the way people do business with the rise of the internet and mobile devices.  People can share their opinions and point of view about business, stocks, events, companies, lifestyle, finance indicators, etc., without the traditional broadcast dependency of radio, television or newspapers.

Networking sites such as Facebook, Twitter, and LinkedIn compete with user's content generators platforms like YouTube and Flickr as well as browsers and engine machines like Google, Yahoo, duckduckgo. These are part of a huge and changing ecosystem of connective media, Van Dijck & and Poell, (2013), Kaplan & Haenlein, (2010), Asur & Huberman, (2010) and Howard & Parks, (2012).

That ecosystem of connective media reflects the relationships between users that have access to the internet, according to Leskovec, Huttenlocher, and Kleinberg, (2010), and suggest that blogs, microblogging, search engines, and content generators add a new perspective to the

social life of the internet-aware people. Due to the high pace and fast human interface, they usually reflect positive and negative aspects that often help to describe certain aspects of our society and business. This is especially true for microblogging platforms, like Twitter.

How can we acquire insight into social media intelligence to support any future investment decisions and actions? We will look into some of these aspects in the next section.

## 2.10. Twitter Feed Impact on Financial Markets

Before continuing, it is necessary to introduce Twitter, since we are going to base our study processing data collected from its platform.

Twitter provides a network that connects users to people, information, ideas, news, debate opinions, and discussions about a wide range of topics (Marketwatch, 2019). From a technical point of view, twitter works as a microblogging platform that enables its users to post a small piece of content which could acquire the form of text, pictures, links, or short videos, along with describing current status or opinion within a limit of 140 characters. The platform has a unique language; however, users have developed some expressions to take advantage of the restriction of the 140 characters.

The @ symbol and username is used to name a person or company i.e @NCI in a tweet. Most people use the $Symbol followed by the company name to refer to the stock of that company, i.e., $Nasdaq, which most likely referred to the Nasdaq composite index.

You can retweet (RT) content and news, which can be a good method to build up connections in your community. You can like other user's content by clicking on the heart that appears to the right of the message of the person or company. One of the most important features is the categorized topic with a hashtag '#'. This helps to disseminate the information on the platform while simultaneously helping to organize it. The hashtags target, for many companies, are used to reveal a large range of preferences and tendencies. For default, each Twitter user is considered a micro-blogger (Battisby, 2019).

In this case, as users we expect a reaction, as the tweet is revealing a personal opinion and perception around a particular topic; in this case, the Nasdaq composite index. This effect would happen in seconds. Yazdanifard et al., (2011) explains that these types of microblogging websites fulfil a need for an even more prompt and faster way of communication. Additionally, the huge amount of data generated in Twitter gives companies the opportunity to reform their strategies based on the latest trends and leads followed by the microbloggers.

Microblogging has become a rich source of data resource, given the number of opinions in different aspects shared on the platform. Twitter is undoubtedly a powerful source for collecting opinions and doing sentiment analysis. This type of analysis is relevant to

understand the new business environment in which companies find themselves. This is because the general level of optimism/pessimism in society affects business decision-makers due to emotions, Anchorage (2014), Anwar Hridoy et al. (2015), Ciftci and Ozturk (2015) and Pak and Paroubek (2016).

There are several methodologies to classified twitter feeds. It depends on the type of information that we can extract from the 140 character content on the tweets. Romero, Meeder, and Kleinberg (2011) explore the way that information spreads between online users. Their objective was to prove that information is spread differently among the users in social networking. They isolated certain topics using the Twitter tokens called: hashtag.

This idea was adapted later for Nisar and Yeung (2018) who used the "trending topics" ideas known as Hashtag in Twitter. They then isolated topics that were related to breaking news, such as terrorist events, or political news. They pulled their data using a hashtag identifier called "Tweetcatcher" for identifying specific trending topics. The relevance of their study was the fact that they were able to isolate trending topics in the UK to study the election and the public's sentiment perspective.

Diakopoulos and Shamma, (2010) also used hashtags to narrow their empirical study about the US 2008 Presidential election. The used search API for grounding common tweets. They also captured the public's sentiment towards the US presidential candidates in 2008 Barack Obama and John McCain and monitored the evolution of said sentiment related to events linked to the presidential campaign. Hashtag classification is a useful tool, however, grounded tweets by the polarity content in user's text would have more impact on the behavioural side of finance.

Bollen, Mao and Zeng, (2010) grounded tweets that contain the word Dow Jones index using the tool "OpinionFinder", which extracts the text polarity (positive and negative sentiment) content in a tweet. After which, he used Granger causality to relate those tweets recorded with the Dow Jones. They showed that the mood levels extracted from users' tweets have predictive value.

Nofer and Hinz, (2015) studied the relationship between mood and the stock market, collecting tweets in Germany. They gained access to the tweets using Twitter API, but they only collect tweetsaccording to the German profile of mood states dictionary or *''Aktuelle Stimmungsskala'' (*ASTS) which is the German profile Mood states integrated by 19 word adjectives, adding all the tweets marked as positive and dividing them among the negatives. They would get a "Social Mood Index", more accurate and easier for analysing.

Other researchers used specific words for analysing, extracting social mood and also market information, as in Porshnev, Redkin and Shevchenko (2013), Azar and Lo (2016) and Reeves (2016). Academics have developed special criteria such as " Retweet count, IsFavorited, User ID, Friends count, Favourites count, Followers count, Tweet text, User language, Tweet

sentiment, and Tweet verb" to select their sampling raw Twitter data, also through Twitter APIs, according to Ranganathan et al (2018).

As we have described, there are different ways for selecting Twitter tweets and extracting sentiment from them. Either using Twitter API, or other software, the main objective is exploring the networking interaction behaviour among users and how this could impact our business.

When it comes to the stock market, most of the researchers have put special attention on the link between the asset value and the social mood. There has been little analysis between the stock market volume and the social mood.

## 2.11. The Psychological Relevance of the Social Mood in Finance

Finance has been built on the traditional precepts of the economy. Fundamentals about the free market, consumer rationality, perfect reflection of the information in the stock price among other arguments were established by Adam Smith and widely used in economics. They are also the foundation of the main capital market theories such as EHM. However, the stock market is also a human interaction networking where participant's opinions and action fluctuates have a large influence on it. Researchers can model stock market behavior on a high fundamental economic perspective but what the participants think about economic fundamentals is what drive their decision-making process, Anchorage (2014).

Visceral factors, according to Loewenstrin (2000), refer to a wide range of negative emotions (anger, fear), drive states (hunger, thirst, sexual desire), and feeling states (pain), that grab people's attention and motivate them to engage in specific behaviours.

When we are evaluating the relationship of social mood and the Nasdaq composite index volume, we are not looking into the causes that can conduct those behavior but rather the effect of them in the stock market, separate to the traditional fundamental or the EHM schools.

Investors and traders, just like people in general, revert to general public behavior due to uncertainty factors, they "follow the crowd", Keynes (1960). This concept is describing as herding. Is is the individual's tendency of following common group behavior instead of deciding based on their own judgment, Baddeley (2010). A few years ago, we did not have enough tools to explore this kind of phenomenon. When the sum of individual pessimism originates "social panic," this will echo the interconnection between risk, anxiety, and fear. This can alter the equilibrium between visceral reactions and cognitive evaluations, Loewenstein et al (2007).

In Chapter 3, we will develop our methodology to test our hypothesis on the link between social mood and the Nasdaq Composite volume Index.

# Chapter 3

# 3. Research Methodology and Method

## 3.1. Introduction

In this section, we will present the two different data resources, from where we retrieved the raw number of positive and negative Tweets, and also the Nasdaq Composite Index historical volume data. We will explain the different methodology for the data analysis process used to answer our dissertation question **"Can Twitter feed analysis assist and ultimately predict an existing correlation between online social mood changes and the daily traded volume on the Nasdaq?**

## 3.2. Method of sampling

We would use two different methods to select the significant number of tweets to test our hypothesis. One from alternative financial data provider pyschosignal.com and the other method using Python's API to mine tweets fed directly from the Twitter platform.

## 3.3. The First Sampling Method

We will get the positive Tweets toward the Nasdaq Composite Index directly from PsychSignal.com.

PsychSignal.com is an innovative social data and sentiment analysis API interactive platform, which enables investors to support their investment decision through additional information resources, like the stock sentiment analysis through Twitter's feeds. The platform operates in partnerships with Twitter and Stocktwits.com. Pyschsignal offer innovative, flexible financial data cloud platforms such as IEXcloud.com, which provides developers and financial platforms with more accessibility and an easier form of integrating into their applications.

The company also offers another feature relating to new sentiment analysis and event reactions. We didn't consider this type of data for this dissertation.

### 3.4. The Second Sampling Method

To extract the raw data to examine the second hypothesis in this dissertation, we created a Twitter developer app to get the necessary key access, for connecting our algorithms to Twitter's Platform.

The keys are an important aspect as they ground the connectivity between Python's library API "Tweepy" and the Twitter platform. We also use a high-level programming language called "Python" for writing our algorithm, with the Python version 2.6, which is a bit old, but it still works quite well. Additionally, we use the Pycharm development environment to do coding using Python. You can find the whole code in Appendix 1. However, we would describe the most important part of the code to set clarity in how it works and for the output we get from it. Note that there is a lot of versions of this methodology available.

There is a different version of this algorithm from a different resource, as the techniques are being widely used currently for enthusiasts' and developers from around the world. We guide our coding following only few resources as (Github, 2019), (LucidProgramming, 2018).

### 3.5. Different API libraries

Using Pycharm, we import the necessary dependencies such as TextBlob, Tweepy, sys, and Matplotlib. We then set the Twitter developer keys in the code, creating the class.

### 3.6. Searching for tweets

This section allows us to put any word that we want, to extract the information and the number of tweets. If we write the word $NASDAQ using the symbol "$" and put the 400, the code would select 400 tweets and extract those tweets containing that specification. See *Figure 3. 1* below.

*Figure 3. 1 Algorithm search and tweets key word input*

```
# input for term to be searched and how many tweets to search
searchTerm = input("Enter Keyword/Tag to search about: ")
NoOfTerms = int(input("Enter how many tweets to search: "))
```

### 3.7. Searching time frame, CSV file, and variable creation

This piece of the code allows us to search tweets from a specific time frame, convert the variable extracted (Tweets), and classify the polarity of the tweet into another variable, such as positive, weakly positive, strongly positive, negative, weakly negative, strongly negative, neutral 8 sub-classes, all together. The result will be shown in a single comma-separated value CSV file. This allows us to store the information extracted and import or export to another file. See *Figure 3. 2* below.

*Figure 3. 2 searching time frame and tweets polarity*

```
# searching for tweets
self.tweets = tweepy.Cursor(api.search, q=searchTerm, since="2019-08-12",
result_type="recent",
                    lang="en").items(NoOfTerms)

# Open/create a file to append data to
csvFile = open('result.csv', 'a')

# Use csv writer
csvWriter = csv.writer(csvFile)

# creating some variables to store info
polarity = 0
positive = 0
wpositive = 0
spositive = 0
negative = 0
wnegative = 0
snegative = 0
neutral = 0
```

### 3.8. Text blob specification.

This section of the algorithm is relating to sentiment text processing. It is a substantial fragment of our code, which tweets sentiment through Textblob, the Python's library language processing (Loria, 2018). This tool is based on common natural language processing NLP. Hence, we can extract the sentiment polarity contained in the tweets.

The polarity is measured in a number range between -1 to 1. The benefit with the API for the learning machine is that, it allows developers to modify most of the functionality, adapting them to specific requirements. Therefore, we adapted the coding to classify the Tweets according to how strong the parity was. i.e., if the polarity of Tweets was greater than 0 but less than 0.3. All the tweets in that range would be weakly positive. We can see all the class criteria below in *Figure 3. 3.*

*Figure 3. 3 Tweets polarity criteria*

```
# iterating through tweets fetched
for tweet in self.tweets:
    #Append to temp so that we can store in CSV later. I use encode UTF-8
    self.tweetText.append(self.cleanTweet(tweet.text).encode('utf-8'))
    # print (tweet.text.translate(non_bmp_map))    #print tweet's text
    analysis = TextBlob(tweet.text)
    # print(analysis.sentiment)  # print tweet's polarity
    polarity += analysis.sentiment.polarity  # adding up polarities to find the average later

    if (analysis.sentiment.polarity == 0):  # adding reaction of how people are reacting to
find average later
        neutral += 1
    elif (analysis.sentiment.polarity > 0 and analysis.sentiment.polarity <= 0.3):
        wpositive += 1
    elif (analysis.sentiment.polarity > 0.3 and analysis.sentiment.polarity <= 0.6):
        positive += 1
    elif (analysis.sentiment.polarity > 0.6 and analysis.sentiment.polarity <= 1):
        spositive += 1
    elif (analysis.sentiment.polarity > -0.3 and analysis.sentiment.polarity <= 0):
        wnegative += 1
    elif (analysis.sentiment.polarity > -0.6 and analysis.sentiment.polarity <= -0.3):
        negative += 1
    elif (analysis.sentiment.polarity > -1 and analysis.sentiment.polarity <= -0.6):
```

It is important to point out that all the tweets analysed were written only in English. It will be interesting to explore the research using API, that pulls and examines data, regardless of the language. In the literature, Pak and Paroubek, (2016) suggest a method based on the special characters contained in the tweets, as the language is not taking into consideration a difference in the text.

### 3.9. Clearing process

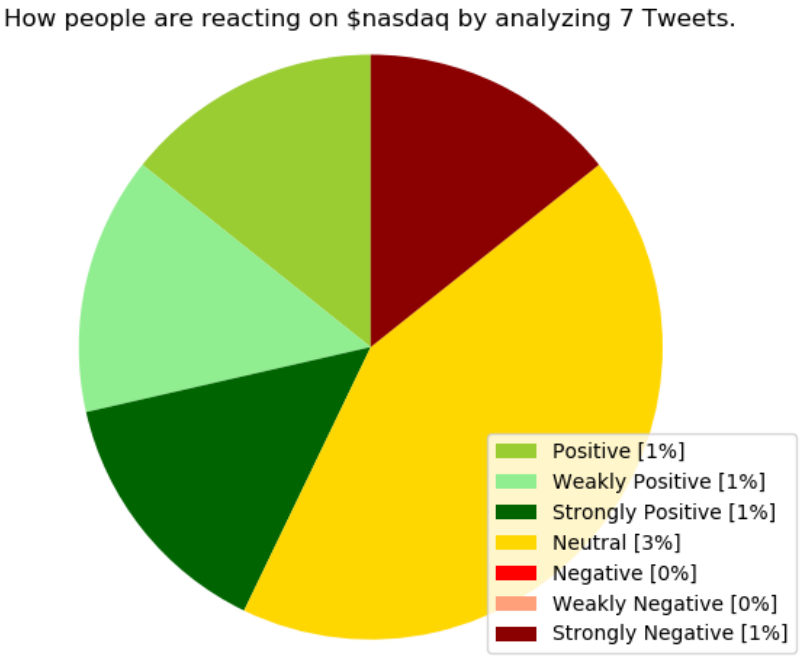
We removed this kind of special character (("(@[A-Za-z0-9]+)|([^0-9A-Za-z \t]) | (\w +:\ / \ / \S +)", " ", ) from our code, as we can see in the piece of code below.

*Figure 3. 4* below displays the outcome when we run the algorithm.

*Figure 3. 4 Tweets cleaning specification*

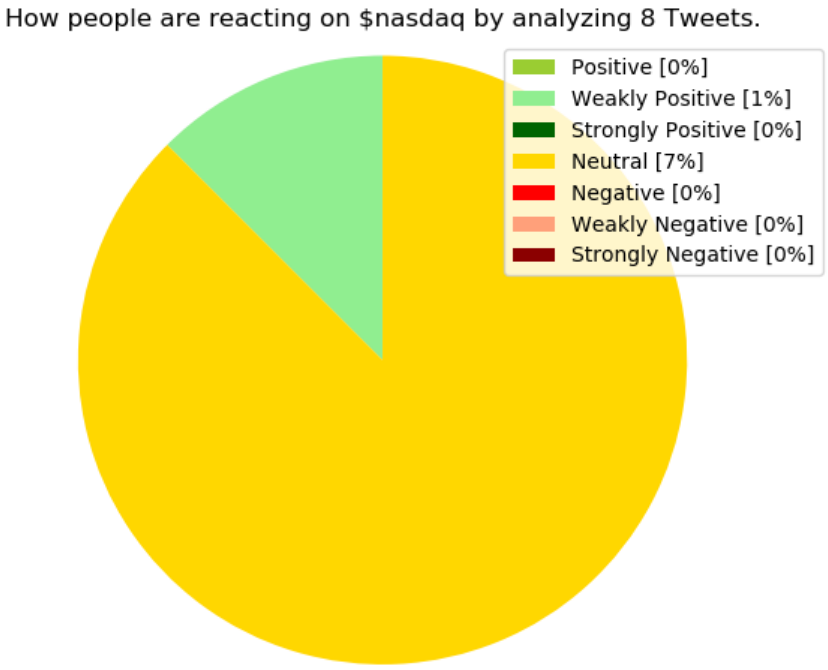
```
def cleanTweet(self, tweet):
    # Remove Links, Special Characters etc from tweet
    return ' '.join(re.sub("(@[A-Za-z0-9]+)|([^0-9A-Za-z \t]) | (\w +:\ / \ / \S +)", " ",
tweet).split())
```

*Figure 3. 5  $nasdaq Twitter sentiment 15 August 2019*




As you can see, we got seven tweets, with the word we target $Nasdaq. The symbol "$" is very important because, usually, people use it along with the asset or company name when they write tweets regarding that particular stock.  We obtained one positive tweet, which polarity was greater than 0.3 and less than 0.6, according to our criteria.  Additionally, to the other tweets that meet the polarity criteria.  The graph is presented as a percentage. However, it is not percentage that is a nominal value.  Each call would be displayed in a CSV chart, showing the outcome according to our criteria and the number of tweets available. You can see that *Figure 3. 5* above shows how the information is displayed in a pie chart when we pull the data using the algorithm. *Figure 3. 6* below also shows an example of how the data is presented in a pie chart, when the algorithm doesn't find some of the criteria that we have previously set up. We got few of them as the data extracted is random and it depend on the number of tweets selected in the moment that we do the call.

*Figure 3. 6 $NASDAQ Twitter sentiment 14 August 2019*




In the Next section below, we would discuss our dataset characteristics, some limitations relating with the Twitter API documentation, along with definitions of our variables to test our model.

### 3.10. Data API Rate Limit

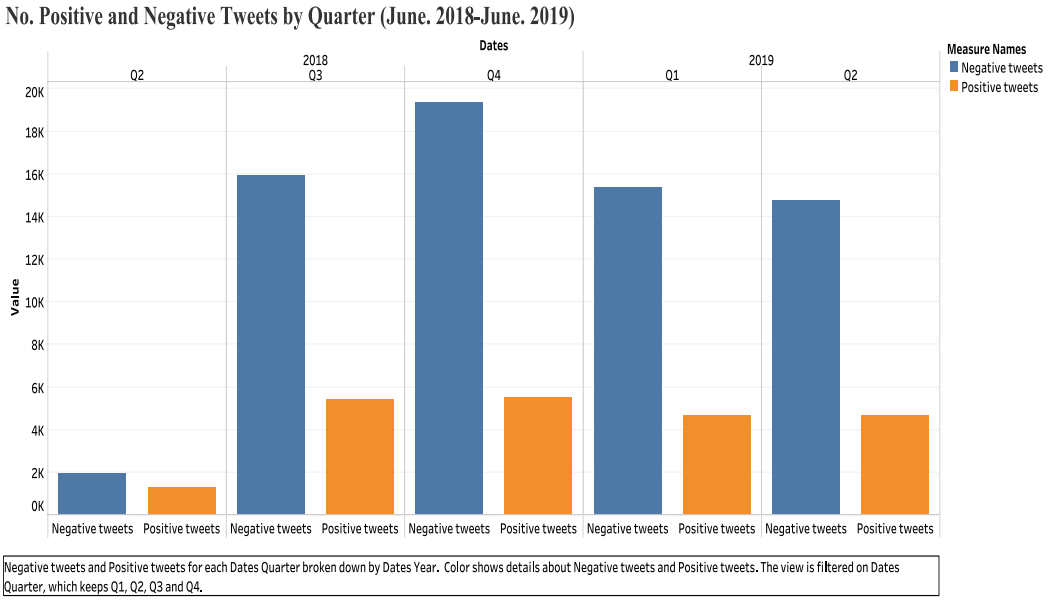
Our process is subject to the Twitter API rate limit page details. The search API is 180 requests per 15 mins window for per-user authentication key. However, we couldn't mine a significant number of tweets as the neutral tweets, where in most cases, were the higher number.

### 3.11. Data organization for the first sampling obtain from pyschosignal.com

We will describe the data obtained from the pyschosignal.com data provider website. We will also define our variables and the main test that we would use to examine the correlation between the variables, under the study described above in Chapter 1.

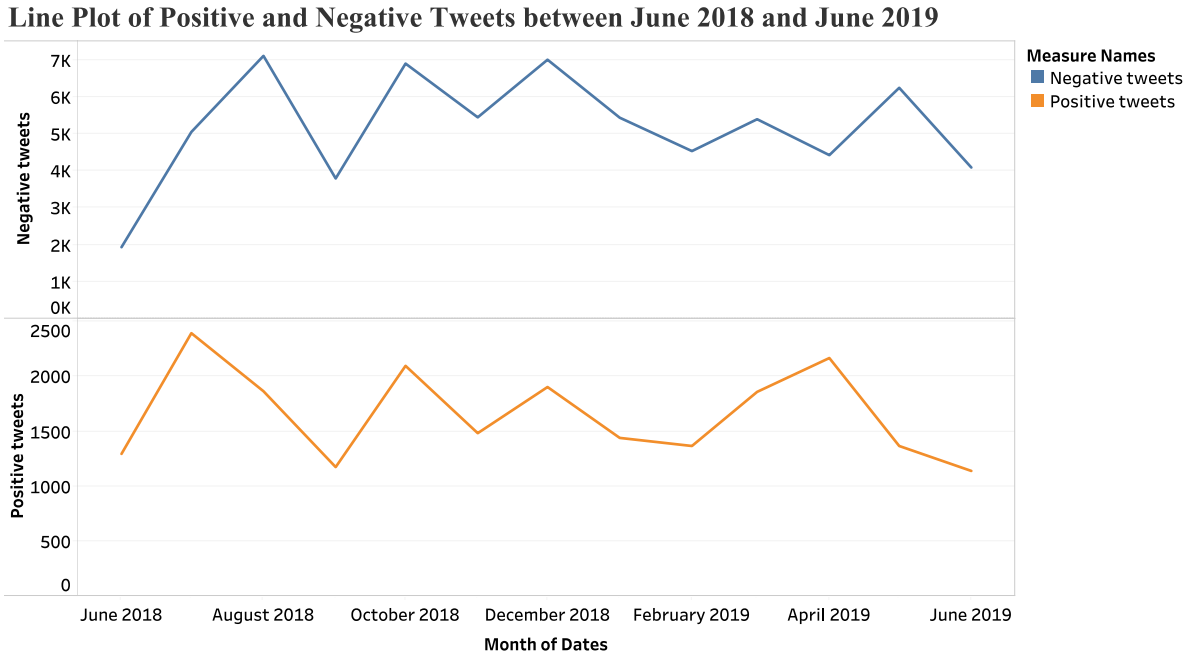
*Figure 3. 7* displays the number of positive and negative tweets by quarter June 2018 to June 2019. It represents the amount of positive and negative Toward the Nasdaq Composite Index that we collect from the Pyschsignal.com.

*Figure 3. 7: Positive and negative tweets by quarter (source:(psychsignal.com, 2019))*




The distribution above shows the sample number of positive and negative tweets that we would use as raw material to analyse our hypothesis. The total volume of negative tweets is higher than the positive tweets in all the quarter. *Figure 3. 8* below shows another perspective of the data described previously.

*Figure 3. 8 The change in the number of positive and negative tweets from June 2018 to June 2019 (source:(psychsignal.com, 2019)*




### 3.11.1 Volume as the Magnitude of Analysis

We follow the techniques applied by (Nisar and Yeung, 2018), to test whether the volume of daily tweets (DVTWS) and the Daily Nasdaq Composite Index (VNIXIC) volume are correlated. A statistical correlations test for exploring the degree of correlation between both variables is used.

### 3.11.2 Independent variable:

The volume of daily tweets = VTWS

The daily Volume of Positive tweets = VPTWS

The daily volume of negative tweets = VNTWS

The *Table 3. 1* displays an example of the number of tweets volume collected. It is important to bear in mind that we have only collected data from one year, but we will not take into account the weekend and some days where there is no official activity At the New York Stock Exchange. This would reduce our observation number to 248.

*Table 3. 1 Twitter volume sentiment indicators.*

| *Dates* | *Total tweets per* | *Positive tweets* | *Negative tweets* |
|---|---|---|---|
| **27/06/2018** | 1913 | 847 | 1066 |
| **28/06/2018** | 867 | 324 | 543 |
| **29/06/2018** | 313 | 78 | 235 |
| **30/06/2018** | 135 | 45 | 90 |
| **01/07/2018** | 183 | 43 | 140 |
| **02/07/2018** | 329 | 103 | 226 |
| **03/07/2018** | 196 | 55 | 141 |
| **04/07/2018** | 81 | 24 | 57 |
| **05/07/2018** | 251 | 66 | 185 |
| **06/07/2018** | 271 | 101 | 170 |
| **07/07/2018** | 104 | 55 | 49 |
| **08/07/2018** | 101 | 36 | 65 |
| **09/07/2018** | 208 | 65 | 143 |
| **10/07/2018** | 376 | 126 | 250 |
| **11/07/2018** | 306 | 120 | 186 |
| **12/07/2018** | 322 | 167 | 155 |

### 3.11.3 Data Limitations

- The number of tweets obtained from the website pyscosignal.com is only from one-year of Twitter sentiment data. To obtain this type of historical data, it is quite expensive, and most of the data providers don't provide their services to a student, only to a company that is officially established.
- As we are studying the effect of daily mood in the Nasdaq Composite (IXIC) volume, the weekend has been filtered for the historical data as there is no trading during that period. Hence, we got a total of 248 observations.

- **Dependent variable:**

The Daily Nasdaq Composite Index volume = VNIXIC

The *Table 3. 2* below display the Nasdaq Composite Volume during the official trading days.

*Table 3. 2 The Daily Nasdaq Composite Index volume indicators*

| Date | Volume |
|---|---|
| **26/06/2018** | 2058640000 |
| **27/06/2018** | 2306430000 |
| **28/06/2018** | 2195400000 |

| | |
|---|---|
| **29/06/2018** | 2192010000 |
| **02/07/2018** | 1767910000 |
| **03/07/2018** | 1179310000 |
| **05/07/2018** | 1745030000 |
| **06/07/2018** | 1704580000 |
| **09/07/2018** | 1837330000 |
| **10/07/2018** | 1725210000 |
| **11/07/2018** | 1761420000 |
| **12/07/2018** | 1926110000 |

Before we set the test, each variable must be normalized using the normalization formula. *Formula, (1),* below, shows the normalized process.

$$\boldsymbol{zi} = \frac{\boldsymbol{xi}-\boldsymbol{\mu}(\boldsymbol{x})}{\boldsymbol{\sigma}(\boldsymbol{x})} \qquad (1)$$

**Where:**

$zi$ = the z-score of x dataset

$\mu(x)$= Mean

$\sigma(x) =$ standard deviation

### 3.11.4 Independent Variable: Twitter Mood

To explore the influence of the independent variable, the total volume of daily 'VTWS', the daily Volume of Positive tweets 'VPTWS,' and the daily volume of negative tweets VNTWS, on the Nasdaq Composite Index volume, the 'Twitter mood' 'TMOOD', the neutral tweets will not be considered in the analysis. *Formula, (2),* shows below the Twitter Mood.

$$\mathbf{TMOOD}_{\boldsymbol{t}} = \frac{(\mathbf{VPTWS}_{\boldsymbol{t}}-\mathbf{VNTWS}_{\boldsymbol{t}})}{\boldsymbol{VTWS}_{\boldsymbol{t}}} \qquad (2)$$

**Where:**

The $\mathrm{TMOOD}_t$ value Could be a range of number between – 1 and 1.

$\mathrm{TMOOD}_t$ would be measured on the same scale than the Textblob Python library sentiment output.

Where 1 represents 100%, and 0 represents neutral tweets. As mentioned before, neutral tweets are those where there are mixed content of positive and negative tweets. Overall, in the study that we have reviewed, there is a high percentage of these type of neutral tweets, proving it

would be interesting to explore a tool to decrease that sort of tweets and incorporate them into the analysis.

### 3.11.5 Dependent Variable: Nasdaq Composite Index Volume

The Nasdaq Composite Index (IXIC) Volume variation, 'VNIXICCHANGE', is established as variation in the volume from the previous day.

*Formula, (3),* shows below the variation in the IXIC

Where:

$$\mathbf{VNIXIC} = \frac{\mathbf{VNIXICCHANGE}_t - \mathbf{VNIXICCHANGE}_{t-1}}{\mathbf{VNIXICCHANGE}_{t-1}} \qquad \mathbf{(3)}$$

### 3.11.6 Regression Measure

For the purpose to test our hypothesis, we will use regression, using the variable described above and we will use the VNIXICCHANGE, VTWS, VPTWS, VNTWS as the input for the formula.

In order to calculate the regression and correlation. *Formula, (4),* shows below the regression.

$$\mathbf{VNIXICCHANGE} = \boldsymbol{a} + \boldsymbol{\beta 1 VPTWS} + \boldsymbol{\beta 2 VNTWS} + \boldsymbol{\beta 3 VTWS} + \boldsymbol{\varepsilon_t} \quad (3)$$

**Where:**

$\varepsilon_t$= random error for the day t
$a$= is the intercept
$\beta$= Are the coefficient for the independent variable Twitter sentiment volume.

### 3.12. Data organization for the second sampling mined using the algorithm in Python

In this section, we will present the raw data that we have mined for the short-window analysis. We establish to do three called a day each day. You can see the result in *Table 3. 3* below.

*Table 3. 3 Number of daily sentiment tweets toward the Nasdaq Index (IXIC).*

| Dates | Positive | weakly positive | Strongly positive | Neutral | Negative | Weakly Negative | Strongly Negative | Total Tweets |
|---|---|---|---|---|---|---|---|---|
| **06/08/2019** | 36 | 25 | 3 | 92 | 8 | 25 | 0 | 189 |
| **07/08/2019** | 25 | 30 | 10 | 99 | 2 | 53 | 3 | 222 |
| **08/08/2019** | 17 | 12 | 8 | 47 | 2 | 15 | 1 | 102 |
| **09/08/2019** | 10 | 11 | 3 | 42 | 0 | 14 | 0 | 80 |

This information would be tested statistically on a short-window analysis, along with the Nasdaq Composite (IXIC), daily volume between August 07$^{th}$ to August 09$^{th,}$ 2019. *Table 3. 4* shows the daily Nasdaq Composite (IXIC) volume to be used to analyse this hypothesis.

*Table 3. 4 The daily Nasdaq Composite (IXIC) volume between August 06$^{th}$ to August 09$^{th,}$ 2019.*

| Dates | Volume |
|---|---|
| **06/08/2019** | 2,2016,100,000 |
| **07/08/2019** | 2,224,330,000 |
| **08/08/2019** | 2,415,670,000 |
| **09/08/2019** | 2,453,230,000 |

# Chapter 4

## 4. Findings and Analysis

### 4.1. Introduction

This dissertation aims to make inferential analysis reading to the relationship between the volume of daily tweets (VTWS), the daily volume of positive tweets (VPTWS) and negative ones (VNTWS), along with the volume of the Nasdaq Composite Index (IXIC). In short, this chapter presents the different findings regarding with the correlation measure between the aforementioned variables. Additionally, the regression and ANOVA results are also presented in this chapter. The same information regarding the relation with the social MOOD and the Volume change in the Nasdaq Composite Index is also presented in this section.

### 4.2. Results first hypothesis

.

*Table 4. 1 the table below shows the first hypothesis that we aim to prove*

| | |
|---|---|
| $Ha1$ **=Hypothesis 1** | There is a significant correlation between Twitter Sentiment towards the Nasdaq Composite Index and the Nasdaq Index Volume for a given day. |
| $H01$**= Null Hypothesis 1:** | There is no significant correlation between Twitter Sentiment towards the Nasdaq Composite Index and the Nasdaq Index Volume for a given day. |

#### 4.2.1. Z-scored ANOVA, regression, and correlation test table results for VTWS, VPTWS, VNWS, VNTWS, VNIXIC.

The regressions test for this and the other test were made on the assumption that the datasets selected are normally distributed. *Table 4.2* below contains the ANOVA outcome for the variance analysis, between the variable described above.

*Table 4. 2: Z-scored ANOVA for VTWS, VPTWS, VNWS, VNTWS, VNIXIC*

| ANOVA[a] | | | | | | |
|---|---|---|---|---|---|---|
| Model | | Sum of Squares | df | Mean Square | F | Sig. |
| 1 | Regression | 21.41 | 2 | 10.705 | 11.626 | .000[b] |
| | Residual | 225.59 | 245 | 0.921 | | |
| | Total | 247 | 247 | | | |
| a. Dependent Variable: Zscore(Volume) | | | | | | |

The next table 4.3 below refer to the Model summary for the Z-scored for the Volume of positive tweets, the volume of negative tweets, Total number of tweets, and the Nasdaq Composite Volume.

*Table 4. 3: Z-scored model summary Regression output for VTWS, VPTWS, VNWS, VNTWS, VNIXIC*

| Model Summary[b] | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate | Change Statistics | | | |
| | | | | | R Square Change | F Change | df1 | df2 |
| 1 | .294[a] | 0.087 | 0.079 | 0.95957 | 0.087 | 11.626 | 2 | 245 |
| a. Predictors: (Constant), Zscore: Total tweets per day, Zscore: Positive tweets | | | | | | | | |
| b. Dependent Variable: Zscore(Volume) | | | | | | | | |

The next table 4.4 content the correlation between the variables under study to support our first hypothesis.

*Table 4. 4: Z-scored correlation coefficient for VTWS, VPTWS, VNWS, VNTWS, VNIXIC*

| Correlations | | | | | |
|---|---|---|---|---|---|
| | | Zscore(Volume) | Zscore: Positive tweets | Zscore: Negative tweets | Zscore: Total tweets |
| Zscore(Volume) | Pearson Correlation | 1 | 0.015 | .245** | .168** |
| | Sig. (2-tailed) | | 0.81 | 0 | 0.008 |
| | N | 248 | 248 | 248 | 248 |
| Zscore: Positive tweets | Pearson Correlation | 0.015 | 1 | .598** | .850** |
| | Sig. (2-tailed) | 0.81 | | 0 | 0 |
| | N | 248 | 248 | 248 | 248 |
| Zscore: Negative tweets | Pearson Correlation | .245** | .598** | 1 | .930** |
| | Sig. (2-tailed) | 0 | 0 | | 0 |
| | N | 248 | 248 | 248 | 248 |
| Zscore: Total tweets | Pearson Correlation | .168** | .850** | .930** | 1 |
| | Sig. (2-tailed) | 0.008 | 0 | 0 | |
| | N | 248 | 248 | 248 | 248 |

### 4.2.2. ANOVA, regression, and correlation test table results for VTWS, VPTWS, VNWS, VNTWS, VNIXIC.

You can find the rest of the statistical set focus on the relation between the IXIC Volume as a dependent variable and the Twitter different polarity as an independent variable.

Table 4.2.3.1: Correlation coefficient for VTWS, VPTWS, VNWS, VNTWS, VNIXIC

| Correlations | | | | | |
|---|---|---|---|---|---|
| | | Volume | Positive tweets | Negative tweets | Total tweets |
| Volume | Pearson Correlation | 1 | 0.015 | .245** | .168** |
| | Sig. (2-tailed) | | 0.81 | 0 | 0.008 |
| | N | 248 | 248 | 248 | 248 |
| Positive tweets | Pearson Correlation | 0.015 | 1 | .598** | .850** |
| | Sig. (2-tailed) | 0.81 | | 0 | 0 |
| | N | 248 | 248 | 248 | 248 |
| Negative tweets | Pearson Correlation | .245** | .598** | 1 | .930** |
| | Sig. (2-tailed) | 0 | 0 | | 0 |
| | N | 248 | 248 | 248 | 248 |
| Total tweets | Pearson Correlation | .168** | .850** | .930** | 1 |
| | Sig. (2-tailed) | 0.008 | 0 | 0 | |
| | N | 248 | 248 | 248 | 248 |
| **. Correlation is significant at the 0.01 level (2-tailed). | | | | | |

The table4.2.3.2 summarized the regression model for the variable's links. Additionally, you can find the summarized test about the population means ANOVA in Table 4.2.3.3 also.

Table 4.2.3.2: Regression output for VTWS, VPTWS, VNWS, VNTWS, VNIXIC

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate | Change Statistics | | |
|---|---|---|---|---|---|---|---|
| | | | | | R Square Change | F Change | df1 |
| 1 | .294[a] | 0.087 | 0.079 | 359016914 | 0.087 | 11.626 | 2 |
| a. Predictors: (Constant), Total tweets per day, Positive tweets | | | | | | | |
| b. Dependent Variable: Volume | | | | | | | |

Table 4.2.3.3: ANOVA for VTWS, VPTWS, VNWS, VNTWS, VNIXIC

| ANOVA[a] | | | | | |
|---|---|---|---|---|---|
| Model 1 | Sum of Squares | df | Mean Square | F | Sig. |
| Regression | 2.99708E+18 | 2 | 1.499E+18 | 11.626 | .000[b] |
| Residual | 3.15788E+19 | 245 | 1.289E+17 | | |
| Total | 3.45759E+19 | 247 | | | |
| a. Dependent Variable: Volume | | | | | |
| b. Predictors: (Constant), Total tweets per day, Positive tweets | | | | | |
| a. Dependent Variable: Volume | | | | | |

In the next Correlation table 4.2.4, contains the correlation result between the MOOD and the Change in the Nasdaq Composite Index volume.

Table 4.2.4: Correlation test table results for MOOD vs. change in the Nasdaq Composite Volume

**Correlations**

| | | MOOD | VNIXIC |
|---|---|---|---|
| MOOD | Pearson Correlation | 1 | .014 |
| | Sig. (2-tailed) | | .832 |
| | N | 249 | 248 |
| VNIXIC | Pearson Correlation | .014 | 1 |
| | Sig. (2-tailed) | .832 | |
| | N | 248 | 248 |

*Table 4.2.5: Model summary table Social Mood (MOOD) & Nasdaq Composite volume change (VNIXIC)*

**Model Summary**

| R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|
| .014 | .000 | -.004 | .184 |

The independent variable is MOOD.

The ANOVA test table 4.2.6. Which gives us an important empirical description right below.

***Table 4.2.6: ANOVA Social Mood & Nasdaq Composite volume change (VNIXIC)***

**ANOVA**

| | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Regression | .002 | 1 | .002 | .045 | .832 |
| Residual | 8.348 | 246 | .034 | | |
| Total | 8.350 | 247 | | | |

The independent variable is MOOD.

### 4.3. Results for the Second Hypothesis

In order to explore whether daily Twitter sentiment volume is a good predictor of the IXIC volume movement, we have to establish a lag series test to determine if the IXIC volume shifts as a consequence of the Twitter sentiment volume. *Table 4. 8* below displays the different correlations and summary model summary for this purpose.

*Table 4.6 Second Hypothesis*

| | |
|---|---|
| $\boldsymbol{Ha2}$ **=Hypothesis 2** | By modeling short daily Twitter sentiment volume and stock volume for the Nasdaq, next day prediction on Nasdaq direction can be made. |
| $\boldsymbol{H02}$ **=Null Hypothesis 2** | By modeling short daily Twitter sentiment volume and stock volume for the Nasdaq, next day prediction on Nasdaq direction can not be made. |

Table 4.8 .1 ANOVA shows the modeling ANOVA for the sample days extracted with the Python API

*Table 4.8.1: ANOVA Social Mood & Nasdaq Composite volume change (VNIXIC)*

**ANOVA[a]**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | .031 | 2 | .016 | 5.497 | .289[b] |
| | Residual | .003 | 1 | .003 | | |
| | Total | .034 | 3 | | | |

a. Dependent Variable: Change VNIXIC

b. Predictors: (Constant), mood_cub, D1MOOD

You can find the model Summary table 4.8.2 of the One-day lag test and the ANOVA table 4.8.3 for the same modelled day next.

*Table 4.8.2: Model summary Social Mood & Nasdaq Composite volume change (VNIXIC) for a time lag of one day*

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .647[a] | .419 | .128 | .0366319 |

a. Predictors: (Constant), one_day_before

*Table 4.8.3: ANOVA Social Mood & Nasdaq Composite volume change (VNIXIC) for a time lag of one day before.*

**ANOVA[a]**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | .002 | 1 | .002 | 1.441 | .353[b] |
| | Residual | .003 | 2 | .001 | | |
| | Total | .005 | 3 | | | |

a. Dependent Variable: Change VNIXIC

b. Predictors: (Constant), one_day_before

The Model Summary and ANOVA test for two-day lag are showing in the next table 4.8.4 and table 4.8.5.

*Table 4.8. 4: ANOVA Social Mood & Nasdaq Composite volume change (VNIXIC) for a time lag of two days before.*

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .831[a] | .690 | .535 | .0280077 |

a. Predictors: (Constant), two_days_before

*Table 4.8.5: ANOVA Social Mood & Nasdaq Composite volume change (VNIXIC) for a time lag of two days before.*

**ANOVA[a]**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | .003 | 1 | .003 | 4.454 | .169[b] |
| | Residual | .002 | 2 | .001 | | |
| | Total | .005 | 3 | | | |

a. Dependent Variable: Change VNIXIC

b. Predictors: (Constant), two_days_before

You can find the same information in the next two those tables regarding the three days lag, modelling for the daily Twitter sentiment and the Nasdaq Composite Index Volume.

*Table 4.8.6: Model summary Mood & Nasdaq Composite volume change (VNIXIC) for a lag time of two days before*

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .831[a] | .690 | .535 | .0280077 |

a. Predictors: (Constant), two_days_before

*Table 4.8.7: Model summary Social Mood & Nasdaq Composite volume change (VNIXIC) for a lag time of three days before*

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .917[a] | .840 | .760 | .0038054 |

a. Predictors: (Constant), three_days_before

*Table 4.8.8: Model summary Mood & Nasdaq Composite volume change (VNIXIC) for a lag of time of three days before*

**ANOVA[a]**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | .000 | 1 | .000 | 10.521 | .083[b] |
| | Residual | .000 | 2 | .000 | | |
| | Total | .000 | 3 | | | |

a. Dependent Variable: Change VNIXIC

b. Predictors: (Constant), three_days_before

# Chapter 5

## 5. Introduction

In this chapter, we will discuss our findings through the analysis of the statistical outcome of the variables described in the previous chapter. We will explore the findings for the first hypothesis, relating specifically to the volume of Twitter sentiments and the Nasdaq Composite Index (IXIC). To prove the link between these variables, we performed a z-core regression, a Pearson' correlation, and ANOVA test as well. We will do the same process using the data mined from the Twitter platform.

### 5.1. Discussion.

We based our dissertation's exploratory test on the assumption that the social mood affects the stock market, used by Bollen, Mao, and Zeng, (2010), broadly suggesting that the general positive or negative public sentiment towards the stock market would impact the affect trading volume. Regarding our first hypothesis, testing the ANOVA in *Table 4.2*: Z-scored, which tells us, it is a significant model. That is to say that the volume of daily tweets (VTWS), the daily volume of positive tweets (VPTWS) and negative ones (VNTWS) are a good predictor of change, with the volume of the Nasdaq Composite Index (IXIC). We determined that 7.9% percentage of the variance in the Nasdaq Composite INDEX Volume Change is explained by the change in the Volume of tweets that contain the word $Nasdaq.

Furthermore, we reject the null hypothesis as the p-value in the ANOVA Table 4.2 Z-score of the total number of tweets, and the IXIC volume is less than the significant test level of 0.05. We can also see this in the table regression table 4.2.3.2, which suggests that the true population correlation coefficient between these variables is not zero. Hence, 7.9 percentage of all the variability in the Nasdaq Composite Index volume could be explained by the Volume of tweets that content some polarity toward the word $Nasdaq, using a dataset of 88950 tweets.

Despite that our findings are small they are aligned with the (Bollen, Mao, and Zeng, 2010; Cropper, 2011; Wei, Mao, and Wang, 2016) results. Although, the statistics test to capture a different result from (Nisar and Yeung, 2018) academic researches as they find a different relation between the twitter Volume and the price change in the FTSE. Although, they inked their variables to certain events like news, which was not considerate in this empirical investigation. Additionally, we observed in the *Table 4.4* that the correlation p-value between Z-score of negative tweets and the Z-score of the Nasdaq Volume are statistically significant the correlation between these variables are different. However, there is a small correlation of r=0.168 between those variables. However, we didn't observe the expected negative correlation between negative tweets and the IXIC volume.

On the other hand, we establish in chapter 1, which we can try to answer through our empirical and statistical analysis established in Chapter 3. In Table 4.2.4, there is a small correlation r=0.14 between the Social mood (MOOD) and the changes in the Nasdaq Composite Index volume. However, we accepted the null hypothesis, as the p-value 0.832 turns out to be greater than the significant level of 0.05. Hence, we assume based on the evidence that we didn't capture a strong correlation between the variables, using the example of 21546 positive tweets and 67389 negative tweets. So, Twitter's sentiment is not a good predictor of the variation in the Nasdaq Composite Stock value, according to the sample selected. In the model summary table 4.2.5, we concluded that the Adjusted R square is -0.004 or -0.4%. However, as we said, the Social mood does not successfully predict a change in the IXIC volume because, in the ANOVA Table 4.2.6, the p-value=0.832 falls out of the significant level of 0.05.

Regarding the second hypothesis, we have discovered a few interesting aspects of our findings. We established a lag time-series test to determine whether the Twitter sentiment was a good predictor of the Nasdaq Composite volume movement. To test our model, we used only the weakly positive and weakly negative tweets toward the $Nasdaq, which were extracted using Python's API algorithms. However, there was not enough evidence to say that we can predict movement in the IXIC volume using Twitter sentiment analysis, as the three lag times series tables p=value is greater than the level of significance. We found that in using the lag 3 days test to measure the predictability of the social mood,

The one-day lag p= value was 0,35, the lag two days p-value was 0.169, and the lag three days p=value was 0.83. Despite the model fail to explain this relationship, and we accepted the null hypothesis. We could observe that the lower p-value and highest F is the three-day lag of day. we could look into the model and move the variables to develop a better model around the three-day lagged variable.

## 5.2. Dataset consideration

From the beginning, we weren't able to get a large dataset as this alternative dataset is quite hard to find. Looking at the dataset at most of the academics research that we have read such as (Zhang, Fuehres and Gloor, 2011; Pak and Paroubek, 2016; Lachanski and Pav, 2017; Jain *et al.*, 2018), we found that the number of tweets involved in their empirical study, were much larger than ours. We capture evidence that the total number of tweets that have a polarity toward the $Nasdaq correlation exists in the population. But we fail to prove the effectiveness of that correlation in the Nasdaq Composite Index volume. So, we suggest for further academics investigations increase the dataset base by using the Python API algorithms mining tweets for some time of at least 3 or 4 years. It appears to be hard work but for the sake of the knowledge is worth it.

On the other hand, Increase the dataset to improve the model according to the second hypothesis that we tried to explore is indispensable. We couldn't reject our null hypothesis as the test might have had low statistical power; thus, we might have been making a Type II error (1-$\beta$), rejecting the null, when we shouldn't. According to what was learned from our quants

lecture, the better way to increase the empirical study power is increasing the sample quality, and sample size by getting more dataset.

# Chapter 6

## 6. Conclusion

Historical data set turns out to be an important part of this type of analysis. However, this type of data is quite expensive. The mayor recommendation is mining data or a period using the python API code that we have used in this dissertation but doing all the call that the Twitter developers documentation allow. API algorithm can be adapted and manipulate for better tweets searching. Sentiment analysis is an exciting topic as the modern evolution in the learning machine as open a door for the behavioural finance study, tools that academics didn't have before and we have the opportunity of improving it through empirical investigations. The twitter polarity is ongoing feel. Even though we found some evidence that the polarity is correlated with the IXIC volume, we couldn't prove its predictability power. However, this a great opportunity as this dissertation added a humble but vital contribution for the search of alternative methods of analysis.

# Bibliography


& J. van D. and Poell, T. (2013) 'Understanding social media', *Understanding Social Media*, 1(1), pp. 1–161. doi: 10.4135/9781446270189.

Agrawal, S. *et al.* (2018) ' Momentum, Mean-Reversion, and Social Media: Evidence from StockTwits and Twitter ', *The Journal of Portfolio Management*, 44(7), pp. 85–95. doi: 10.3905/jpm.2018.44.7.085.

Anchorage, A. (2014) 'Social Mood and Financial Economics', *Journal of Behavioral Finance*, (April 2003). doi: 10.1207/s15427579jpfm0603.

Anwar Hridoy, S. A. *et al.* (2015) 'Localized twitter opinion mining using sentiment analysis', *Decision Analytics*. Springer Berlin Heidelberg, 2(1). doi: 10.1186/s40165-015-0016-4.

Azar, P. and Lo, A. W. (2016) 'The Wisdom of Twitter Crowds: Predicting Stock Market Reactions to FOMC Meetings via Twitter Feeds', *Ssrn*. doi: 10.2139/ssrn.2756815.

Baddeley, M. (2010) 'Herding, social influence and economic decision-making: Socio-psychological and neuroscientific analyses', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1538), pp. 281–290. doi: 10.1098/rstb.2009.0169.

Battisby, A. (2019) *Twitter overview*. Available at: https://www.marketingdonut.co.uk/social-media/twitter/twitter-overview (Accessed: 26 July 2019).

Beja, A. (1977) 'The Limits of Price Information in Market Processes'.

Bollen, J., Mao, H. and Zeng, X. (2010) 'Twitter mood predicts the stock market .', pp. 1–8.

Broadstock, D. C. and Zhang, D. (2019) 'Social-media and intraday stock returns: The pricing power of sentiment', *Finance Research Letters*. Elsevier, 30(March), pp. 116–123. doi: 10.1016/j.frl.2019.03.030.

Chen, H. *et al.* (2014) 'Wisdom of crowds: The value of stock opinions transmitted through social media', *Review of Financial Studies*, 27(5), pp. 1367–1403. doi: 10.1093/rfs/hhu001.

Ciftci, K. and Ozturk, S. S. (2015) 'A Sentiment Analysis of Twitter Content as a Predictor of Exchange Rate Movements', (November). doi: 10.13140/RG.2.1.1022.9201.

Cropper, A. (2011) 'Modelling Stock Volume Using Twitter', (September).

Damasio, A. (1995) *Damasio, Antonio R., Descartes' Error: Emotion, Reason, and the Human Brain, Relations industrielles*. doi: 10.7202/051028ar.

Darskuviene, V. (2010) 'Financial Markets', *Financial Markets*, p. 140.

Daves, P. R. (2003) 'Reported trading volume on the NYSE and NASDAQ', (May 2003), pp. 1–29.

Diakopoulos, N. A. and Shamma, D. A. (2010) 'Characterizing debate performance via aggregated twitter sentiment', p. 1195. doi: 10.1145/1753326.1753504.

Eugene and 1965, F. (1965) 'The behavior of stock-Market Prices', *Chicago Journals*, 38(1), pp. 34–105.

Fama, B. E. F. and Fama, E. U. F. (1965) 'Random Walks in Stock- Market Prices'.

Garc, B., Nieto, L. and Valencia, S. (2017) ‘“ C Hartist a Nalysis ”’.

Github (2019) *twitter-sentiment-analysis*. Available at: https://github.com/topics/twitter-sentiment-analysis (Accessed: 1 July 2019).

Internet world Stats (2019) *Internet World Stats*. Available at: https://internetworldstats.com/stats.htm.

Jain, A. *et al.* (2018) ‘Forecasting Price of Cryptocurrencies Using Tweets Sentiment Analysis’, *2018 11th International Conference on Contemporary Computing, IC3 2018*, (February 2019), pp. 0–7. doi: 10.1109/IC3.2018.8530659.

Jiranyakul, K. (2007) ‘Dynamic relatioship between stock return, trading volume, and volatility in the stock exchange of Thailand does the US subprime crisis matters’, *Economic Policy*, (2116), pp. 0–33. doi: 10.1227/01.NEU.0000349921.14519.2A.

Karpoff, J. M. (2006) ‘The Relation Between Price Changes and Trading Volume: A Survey’, *The Journal of Financial and Quantitative Analysis*, 22(1), p. 109. doi: 10.2307/2330874.

Kim, J. R. (no date) ‘Measuring the Intrinsic Value Ja Ryong Kim* The University of Edinburgh Business School’, *The University of Edinburgh Business School, 29 Buccleuch Place, Edinburgh EH8 9JS, UK*, 44(0), pp. 0–36.

Kumar, A. (2009) ‘Who Gambles? in the Stock Market’, *Journal of Finance*, 64(4), pp. 1889–1933. doi: 10.1111/j.1540-6261.2009.01483.x.

Kumar, A., Page, J. K. and Spalt, O. G. (2011) ‘Religious beliefs, gambling attitudes, and financial market outcomes’, *Journal of Financial Economics*. doi: 10.1016/j.jfineco.2011.07.001.

Kumar, A., Page, J. K. and Spalt, O. G. (2016) *Gambling and Comovement, Journal of Financial and Quantitative Analysis*. doi: 10.1017/S0022109016000089.

Lachanski, M. and Pav, S. (2017) ‘Shy of the character limit: “Twitter mood predicts the stock market” revisited’, *Econ Journal Watch*, 14(3), pp. 302–345.

Laurell, C. and Sandström, C. (2017) ‘The sharing economy in social media: Analyzing tensions between market and non-market logics’, *Technological Forecasting and Social Change*. Elsevier, 125(June 2016), pp. 58–65. doi: 10.1016/j.techfore.2017.05.038.

Leskovec, J., Huttenlocher, D. and Kleinberg, J. (2010) ‘Signed networks in social media’, *Proceedings of the 28th international conference on Human factors in computing systems - CHI ’10*, p. 1361. doi: 10.1145/1753326.1753532.

Loria, S. (2018) *TextBlob: Simplified Text Processing*. Available at: https://textblob.readthedocs.io/en/dev/.

LucidProgramming (2018) *Twitter API with Python: Part 1 -- Streaming Live Tweets*. Available at: https://www.youtube.com/watch?v=wlnx-7cm4Gg&list=FLb6e2Ifo13W_qRsohCGvkqQ&index=29&t=665s (Accessed: 10 August 2019).

marketwatch (2019) *Twitter Inc.* Available at: https://www.marketwatch.com/investing/stock/twtr (Accessed: 22 July 2019).

Mehmed, L. quang T. and mustafa (2009) ‘the Relationship Between Trading Volume , Stock Index Returns and Volatility ’:, pp. 1–34.

Nasdaq (2019) *Nasdaq Composite*. Available at: https://indexes.nasdaqomx.com/docs/FS_COMP.pdf (Accessed: 30 July 2019).

NASDAQ (2017) *NASDAQ Composite Index® Methodology*. Available at: https://indexes.nasdaqomx.com/docs/Methodology_COMP.pdf (Accessed: 30 July 2019).

Nisar, T. M. and Yeung, M. (2018) 'Twitter as a tool for forecasting stock market movements: A short-window event study', *The Journal of Finance and Data Science*. Elsevier Ltd, 4(2), pp. 101–119. doi: 10.1016/j.jfds.2017.11.002.

Nofer, M. and Hinz, O. (2015) 'Using Twitter to Predict the Stock Market: Where is the Mood Effect?', *Business and Information Systems Engineering*. Springer Fachmedien Wiesbaden, 57(4), pp. 229–242. doi: 10.1007/s12599-015-0390-4.

Pak, A. and Paroubek, P. (2016) 'Twitter as a Corpus for Sentiment Analysis and Opinion Mining', *Ijarcce*, 5(12), pp. 320–322. doi: 10.17148/IJARCCE.2016.51274.

Parracho, P., Neves, R. and Horta, N. (2010) 'Trading in financial markets using pattern recognition optimized by genetic algorithms', p. 2105. doi: 10.1145/1830761.1830884.

Perry, E. (2017) *What exactly IS an API?* Available at: https://medium.com/@perrysetgo/what-exactly-is-an-api-69f36968a41f.

Porshnev, A., Redkin, I. and Shevchenko, A. (2013) 'Improving Prediction of Stock Market Indices by Analyzing the Psychological States of Twitter Users', *Ssrn*. doi: 10.2139/ssrn.2368151.

psychsignal.com (2019) *IXIC Mood*. Available at: https://psychsignal.com/ (Accessed: 1 July 2019).

Ranco, G. *et al.* (2015) 'The effects of twitter sentiment on stock price returns', *PLoS ONE*, 10(9), pp. 1–21. doi: 10.1371/journal.pone.0138441.

Ranganathan, J. *et al.* (2018) 'Actionable pattern discovery for Sentiment Analysis on Twitter Data in clustered environment', *Journal of Intelligent and Fuzzy Systems*, 34(5), pp. 2849–2863. doi: 10.3233/JIFS-169472.

Reeves, T. J. (2016) 'Sentiment Analysis for Long-Term Stock Prediction by Tyler Joseph Reeves A Thesis Presented in Partial Fulfillment of the Requirements for the Degree Master of Science Approved April 2016 by the Graduate Supervisory Committee : Hasan Davulcu , Chair John', (April).

Roberts, A. (2003) 'The efficent market hypothesis and its critics', *Undang-Undang Republik Indonesia Nomor 20 Tahun 2003 Tentang Sistem Pendidikan Nasional Dengan Rahmat Tuhan Yang Maha Esa Presiden Republik Indonesia*, 14(1), pp. 1–26.

Romero, D. M., Meeder, B. and Kleinberg, J. (2011) 'Differences in the mechanics of information diffusion across topics', p. 695. doi: 10.1145/1963405.1963503.

Rout, J. K. *et al.* (2018) 'A model for sentiment and emotion analysis of unstructured social media text', *Electronic Commerce Research*. Springer US, 18(1), pp. 181–199. doi: 10.1007/s10660-017-9257-8.

See-To, E. W. K. and Yang, Y. (2017) 'Market sentiment dispersion and its effects on stock return and volatility', *Electronic Markets*. Electronic Markets, 27(3), pp. 283–296. doi: 10.1007/s12525-017-0254-5.

Sewell, M. (2012) ‘The E ffi cient Market Hypothesis : Empirical Evidence’, 1(2).

Shiller, R. J. (1999) ‘Human Behavior and the Efficiency of Financial Markets’, *Handbook of Macroeconomics Volume 1*, pp. 1305–40.

Twitter (2019) *Twitter developers*. Available at: https://developer.twitter.com/en/docs.html.

Velay, M. and Daniel, F. (2018) ‘Stock Chart Pattern recognition with Deep Learning’. Available at: http://arxiv.org/abs/1808.00418.

Wang, J. C. & S. G. & J. (1993) ‘trading volume and serial correlation in stock returns’.

Wei, W., Mao, Y. and Wang, B. (2016) ‘Twitter volume spikes and stock options pricing’, *Computer Communications*. Elsevier Ltd., 73, pp. 271–281. doi: 10.1016/j.comcom.2015.06.018.

Xiang, Z. and Gretzel, U. (2010) ‘Role of social media in online travel information search’, *Tourism Management*. Elsevier Ltd, 31(2), pp. 179–188. doi: 10.1016/j.tourman.2009.02.016.

Yazdanifard, R. *et al.* (2011) ‘Social Networks and Microblogging ; The Emerging Marketing Trends & Tools of the Twenty-first Century’, *Management*, 5, pp. 577–581.

Zeng, D. *et al.* (2010) ‘Social media analytics and intelligence’, *IEEE Intelligent Systems*. IEEE, 25(6), pp. 13–16. doi: 10.1109/MIS.2010.151.

Zhang, X., Fuehres, H. and Gloor, P. A. (2011) ‘Predicting Stock Market Indicators Through Twitter “I hope it is not as bad as I fear”’, *Procedia - Social and Behavioral Sciences*. Elsevier B.V., 26(2007), pp. 55–62. doi: 10.1016/j.sbspro.2011.10.562.

Ziembinski, B. (2015) ‘Social mood revealed’, *CEUR Workshop Proceedings*, 1351, pp. 35–50.

# Appendices

## APPENDIX 1

API python code.

```
      # input for term to be searched and how many tweets to search
         searchTerm = input("Enter Keyword/Tag to search about: ")
         NoOfTerms = int(input("Enter how many tweets to search: "))

         # searching for tweets
         self.tweets = tweepy.Cursor(api.search, q=searchTerm, since="2019-08-15",
result_type="recent",
                                     lang="en").items(NoOfTerms)

         # Open/create a file to append data to
         csvFile = open('result.csv', 'a')

         # Use csv writer
         csvWriter = csv.writer(csvFile)

         # creating some variables to store info
         polarity = 0
         positive = 0
         wpositive = 0
         spositive = 0
         negative = 0
         wnegative = 0
         snegative = 0
         neutral = 0


         # iterating through tweets fetched
         for tweet in self.tweets:
             #Append to temp so that we can store in csv later. I use encode UTF-8
             self.tweetText.append(self.cleanTweet(tweet.text).encode('utf-8'))
             # print (tweet.text.translate(non_bmp_map))    #print tweet's text
             analysis = TextBlob(tweet.text)
             # print(analysis.sentiment)  # print tweet's polarity
             polarity += analysis.sentiment.polarity  # adding up polarities to
find the average later

             if (analysis.sentiment.polarity == 0):  # adding reaction of how
people are reacting to find average later
                neutral += 1
             elif (analysis.sentiment.polarity > 0 and analysis.sentiment.polarity
<= 0.3):
                wpositive += 1
             elif (analysis.sentiment.polarity > 0.3 and
analysis.sentiment.polarity <= 0.6):
                positive += 1
             elif (analysis.sentiment.polarity > 0.6 and
analysis.sentiment.polarity <= 1):
                spositive += 1
             elif (analysis.sentiment.polarity > -0.3 and
analysis.sentiment.polarity <= 0):
                wnegative += 1
             elif (analysis.sentiment.polarity > -0.6 and
analysis.sentiment.polarity <= -0.3):
                negative += 1
             elif (analysis.sentiment.polarity > -1 and analysis.sentiment.polarity
<= -0.6):
                snegative += 1
```

```
        # Write to csv and close csv file
        csvWriter.writerow(self.tweetText)
        csvFile.close()

        # finding average of how people are reacting
        print("This is the total number of tweets by sectors: ")
        print(positive, wpositive, spositive, negative, wnegative, snegative,
neutral)

        totaltweets = sum([positive, wpositive, spositive, negative, wnegative,
snegative, neutral])
        if (totaltweets != NoOfTerms):
            totalitems = "Number of Tweets for positive {}, wpositive {},
spositive {}, negative {}, wnegative {}, snegative {}, neutral {}".format(
                positive, wpositive, spositive, negative, wnegative, snegative,
neutral)
        print(totalitems)
        NoOfTerms = totaltweets

        # finding average reaction
        polarity = polarity / NoOfTerms

        # printing out data
        print("How people are reacting on " + searchTerm + " by analyzing " +
str(NoOfTerms) + " tweets.")
        print()
        print("General Report: ")

        if (polarity == 0):
           print("Neutral")
        elif (polarity > 0 and polarity <= 0.3):
           print("Weakly Positive")
        elif (polarity > 0.3 and polarity <= 0.6):
           print("Positive")
        elif (polarity > 0.6 and polarity <= 1):
           print("Strongly Positive")
        elif (polarity > -0.3 and polarity <= 0):
           print("Weakly Negative")
        elif (polarity > -0.6 and polarity <= -0.3):
           print("Negative")
        elif (polarity > -1 and polarity <= -0.6):
           print("Strongly Negative")

        print()
        print("Detailed Report: ")
        print(str(positive) + "% people thought it was positive")
        print(str(wpositive) + "% people thought it was weakly positive")
        print(str(spositive) + "% people thought it was strongly positive")
        print(str(negative) + "% people thought it was negative")
        print(str(wnegative) + "% people thought it was weakly negative")
        print(str(snegative) + "% people thought it was strongly negative")
        print(str(neutral) + "% people thought it was neutral")

        self.plotPieChart(positive, wpositive, spositive, negative, wnegative,
snegative, neutral, searchTerm, NoOfTerms)


    def cleanTweet(self, tweet):
        # Remove Links, Special Characters etc from tweet
        return ' '.join(re.sub("(@[A-Za-z0-9]+)|([^0-9A-Za-z \t]) | (\w +:\ / \ /
\S +)", " ", tweet).split())

    # function to calculate percentage
    def percentage(self, part, whole):
        temp = 100 * float(part) / float(whole)
        return format(temp, '.2f')
```

```
    def plotPieChart(self, positive, wpositive, spositive, negative, wnegative,
snegative, neutral, searchTerm, noOfSearchTerms):
        labels = ['Positive [' + str(positive) + '%]', 'Weakly Positive [' +
str(wpositive) + '%]','Strongly Positive [' + str(spositive) + '%]', 'Neutral [' +
str(neutral) + '%]',
                  'Negative [' + str(negative) + '%]', 'Weakly Negative [' +
str(wnegative) + '%]', 'Strongly Negative [' + str(snegative) + '%]']
        sizes = [positive, wpositive, spositive, neutral, negative, wnegative,
snegative]
        colors = ['yellowgreen','lightgreen','darkgreen', 'gold',
'red','lightsalmon','darkred']
        patches, texts = plt.pie(sizes, colors=colors, startangle=90)
           .legend(patches, labels, loc="best")
        plt.title('How people are reacting on ' + searchTerm + ' by analyzing ' +
str(noOfSearchTerms) + ' Tweets.')
        plt.axis('equal')
        plt.tight_layout()
        plt.show()

if __name__ == "__main__":
     sa = SentimentAnalysis()
     sa.DownloadData()

for tweet in self.tweets:
     print(tweet.text)
     analysis = TextBlob(tweet.text)
```