LUND UNIVERSITY
School of Economics and Management

Master's Programme in International Economics with a focus on China

The Forecasting of Daily Stock Returns with Sentiment Analysis and Google Trends

Exploring the Apple and Amazon Case

by

Pavel Zadražil

pa8856za-s@student.lu.se

Abstract: Previous research has shown a link between investor sentiment and stock market movement, as well as the correlation of search engine queries and market volatility. This research aims to extend these findings by predicting the exact firm-level daily return of Apple and Amazon. The daily returns are predicted with an artificial neural network model. In theory, this model is able to learn similarly to a human brain, it finds the relevance of individual inputs (investor sentiment and investor interest to trade) and then computes daily returns based on the inputs' significance. Even though the model used in this research has an ability to learn, it fails to predict exact daily returns. However, the model is able to predict the positive or negative direction of the stock returns fairly accurately.

Keywords: Predict; Equity Market; Daily Returns; Artificial Neural Network; Investor Sentiment; Investor Interest

EKHS51
Master's Thesis (15 credits ECTS)
June 2019
Supervisor: Seán Kenny
Examiner: Anders Ögren

Acknowledgements

I would like to acknowledge my supervisor for his positive and calm attitude. My grandma, brother, friends and all the people who believe in me for their support throughout my studies. Last but not least, to tribute each and every teacher who challenges, respects and loves what she/he does. I would not get this far without you.

Thank you!

Table of Contents

List of Tables

List of Figures

1 Introduction

The internet has changed the way investors investigate equity markets and invest in general. Equity investors used to have to wait for annual reports, printed press, and meet their broker in order to adjust their portfolio. Nowadays, investors are able to manage it with their smartphones. They analyse their portfolios every day and resell stocks via broker applications - it creates a highly dynamic market (Hagstrom, 2013). This change represented in figures; shows that 90 per cent of data in the world was generated in the last 2 years, more than half of the global population has access to the internet, 456,000 tweets are posted every minute of a day and Google engine processes 3.5 billion queries a day (Marr, 2018, A), we live in a data-driven world.

According to Oh and Kim (2002), a stock market is influenced by factors such as political decisions, market conditions, annual reports and investors' expectations. Even though not all the tweets, news and political decision are related to the market situation, it is still impossible to process the entire scale of relevant information since the data volume grows exponentially. Thus, this research aims to forecast short-term stock prices by focusing exclusively on two important quantities: investor sentiment and investor interest.

In theory, all influencing quantities suggested by Oh and Kim (2002) might be reflected in investors sentiment, because, in the end, it always comes down to investor sentiment. The sentiment is the decision-maker which determines if stocks should be purchased, kept, or sold. For instance, an annual report exceeding investors' expectations, results in a positive change of investors sentiment, poor political reform and an announcement of a CEO retirement, lowers their sentiment. However, investor sentiment alone is insufficient, as it is essential to identify their aim of which stock is going to be traded, which, in this thesis is defined as investor interest.

1.1 Research Problem

This thesis uses a new unique way of forecasting market return combining two methods: investor sentiment and investor interest to trade. Furthermore, a comparison will be provided of two proxies for investor sentiment: the first method collects data from Twitter, and the second method uses news. The result should not only demonstrate which proxy is a better predictor, it should also reflect what kind of society we live in. In recent years, the reliability of politics-related news and social media overuse have been challenged, Twitter especially, as a personal micro-blogging channel. The results should uncover whether investors depend on social media or traditional media. Lastly, preceding empirical studies by Bollen, Mao and Zeng's (2011), and Mittal and Goel (2012) of investor sentiment focus on behaviour of the whole market using, for example, Dow Jones Industrial Average (DJIA) as a proxy for market movement. However, this study operates on a company level; respectively it tests the theory on two corporations: Apple Inc. and Amazon.com, Inc.

Since this is a company level study, the investor sentiment must be on the company level; it would not be efficient to use a sentiment which indicates the whole market conditions. The model must be reflecting investor interest to trade Apple or Amazon because the sentiment by itself might not have to be relevant if investors are not willing to trade the stock. In that regard, some theorists (Bordino, Battiston, Caldarelli, Cristelli, Ukkonen & Weber, 2012; Dimpfl & Jank 2016; Stephens-Davidowitz & Pabon, 2017) argue that search engine queries are able to forecast market volatility. Consequently, search engine queries might be a sufficient proxy for investor interest.

1.2 Aim and Scope

This research aims to forecast the daily returns of two companies: Apple and Amazon, which are listed on the NASDAQ stock exchange in the United States of America. The study is conducted throughout 3 years' time frame from January 2015 to December 2018, and it uses daily data. During this time period, modern technologies have generated more than 90 per cent of data in the world (Marr, 2018, A). This data-oriented shift indicates a modern approach in market research spheres. Therefore, this research uses an artificial neural network to process the data and predict daily stock returns. Forecasting of daily returns seem to be a challenging task, scholars often investigate weekly or monthly data instead. Even though, investors' horizons becoming shorter every year and their focus is shifting towards short term resale prices (Lee, Shleifer & Thaler 1991). Lastly, the research question is investigated on two investor sentiments: Twitter and news. It is for comparison purposes and robustness of the investigation.

1.3 Outline of the Thesis

The strategy of this thesis proceeds as follows: next theoretical section identifies suitable proxies for investor sentiment and investor interest based on previous academic studies. Based on the theoretical findings, the research question and 4 hypotheses are introduced. The third section presents sources from where the data was collected and argues its reliability and suitability. Which is followed by section 4 that introduces the artificial neural network model. Section 5 is divided into 2 subsections. Firstly, it replicates and extends Bordino's et al. (2012) research in order to verify the effect of Google search engine data on Apple and Amazon stock volatility; then the thesis results are presented. The final section concludes the findings.

2 Theory

The theory section establishes the theoretical background of this study, evaluates suitable proxies for investor sentiment, investor interest and clarify what drives investors' decision making and daily market volatility. Then it introduces the research question and 4 hypotheses based on previous researches.

2.1 Previous Research

In the 1960s, Fama (1960) has introduced his Efficient Market Hypothesis. This hypothesis suggests intelligent and well-informed investors who adjust equity price based on all available information. However, equity market history has shown several rationally unexplainable events: The Great Crash, Go-Go Years, Black Monday and Dot.com bubble. All these events lack rational explanation within the standard finance model (Baker & Wurgler, 2007). Behavioural economics has a potential explanation. For instance, Hagstrom (2013) describes that investors might behave as a herd, follow a trend without an understanding of the core value of the investment. De Long, Shleifer, Summers and Waldmann (1990) argue that noise trader risk causes market abnormality and volatility. Furthermore, there is no evidence that the actual value of investment changes accordingly to the price changes (Shiller, 1987). It suggests that transcend volatility and market anomaly to some extent is caused by irrational investors or investors with a tendency to overreact who invest accordingly to their sentiment.

2.1.1 Investor Sentiment

"It has long been market folklore that the best time to buy stocks is when individual investors are bearish, and the best time to sell is when individual investors are bullish." by Neal and Wheatley (1998, p. 523). Investor sentiment provides enlightenment into investors' reasoning whether their mindset is bearish or bullish. The issue is which proxy of investor sentiment is the most accurate representation of the investor mindset. The next section covers investor sentiment proxies used in previous studies and argues which is the most suitable measurement for daily stock prediction.

Proxies of Investor Sentiment

It is evident from the literature (Neal & Wheatley, 1998; Baker & Wurgler, 2006; Lemmon & Portniaguina, 2006) that investor sentiment influences an equity market. However, scholars are not in agreement which proxy is the most suitable measurement of the sentiment. For instance, Baker and Wurgler (2006, p. 1655) construct their investor sentiment by combining," the closed-end fund discount, NYSE share turnover, the number and average first-day returns on IPOs, the equity share in new issues, and the dividend premium." Their opinion is partly consistent with Neal and Wheatley (1998) who suggest that the closed-end fund discount, redemptions of net mutual fund and the ratio of odd-lot sales to purchases are applicable investor sentiment measurements. On the other hand, Lemmon and Portniaguina (2006) apply consumer confidence as a proxy of investor sentiment. However, these proxies are not applicable to this study. They are excessively complex and reflect the entire market situation. Although a company-level research requires a focused approach.

The internet, big data respectively, have enabled to measure new metrics. This technological shift is noticeable in marketing. Companies such as Facebook and Google built their wealth and position through control of the result. They have created customers profiles, which enable to target a specific group of people and collect data about them. This is one of the main advantages in comparison to traditional media such as television, radio or newspaper, where the audience profiling is rather limited. In a similar matter, this study requires proxy for investor sentiment, which reflects what is investors current (daily) impression of Apple or Amazon. However, for instance, the number and average first-day returns on IPOs measurement require months of data collecting in order to acquire a non-bias sentiment proxy.

As was mentioned in the introduction, 90 per cent of data in the world was generated in the last 2 years (Marr, 2018, A). A considerable amount is generated online, which makes it impossible for a human to process it. However, computers are able to process extensively large quantum of data. Secondly, computers do not have emotions, they consistently examine the data without any exception. According to Stephens-Davidowitz and Pabon (2017), people do not have the ability. Due to the fact that people process text subjectively, which results in questionnaire bias, for instance, consumer confidence as an investor sentiment proxy is predisposed to be biased. Therefore, recent studies adopt a big data approach, where Natural Language Processing (NLP), a subfield of machine learning which analyses text (Wahome, 2018). The text can be distributed through social media, online news and personal blogs, then investor sentiment will be computed by NLP. One of the main advantages is the size of the online world. For example, there are 330 million active users on Twitter monthly (Wahome, 2018). Investors are influenced by tweets as well as consumers. They both influence future stock return, investors directly by acquiring a company shares and consumers indirectly by adjusting future revenues.

There are several studies researching the relationship between the online world sentiment and equity market returns. Bollen, Mao and Zeng's (2011) paper identifies a correlation between Twitter sentiment and DJIA index. Their paper was to some extent replicated by Mittal and Goel (2012). Both papers show a high percentage correlation between calm sentiment and DJIA, 87 per cent and 75.56 per cent respectively. Moreover, Mittal and Goel (2012) find a correlation for happiness dimension, which has not been found in the original paper. Overall, their researches suggest relationship only for 2 categories out of 6 namely: "Calm", "Alert", "Sure", "Vital", "Kind" and "Happy" computed by Google-Profile of Mood States method. Rao and Srivastava (2012) use different metrics: negative and positive sentiment, their results show strong cause-effect with stock return. In contrary, Bollen, Mao and Zeng's (2011) do not uncover this relationship while applying negative and positive sentiment analysis. This extensive disparity might be due to a different algorithm, which has been used to calculate the sentiment.

Another way of acquiring the sentiment is through the news. Tetlock (2007) has been the first scholar to forecast DJIA returns with news articles, he collected the data from Wall Street Journal columns. He tests for 3 negative sentiments: Pessimism, Negative and Weak. His research suggests that negative news sentiment has a statistically and economically significant impact on an equity market. Mian and Sankaraguruswamy (2012) analyse how company earnings reported in news affect stock market movement. Their research indicates persistent stock price movement up to 60 days after announcing earnings. Their results are statistically significant and consistent with the literature. It indicates that announced earnings influence stock prices.

The literature proposes several alternatives of sentiments that might be a suitable proxy for investor sentiment. Unlike "Kind", "Alert", "Sure" and "Vital" classifications which do not provide information about stock movements (Bollen, Mao & Zeng, 2011; Mittal & Goel, 2012). This thesis uses Twitter and news sentiment which appears to be the most suitable for the research of an individual company. Due to its ability to mine data of a specific company. Besides, the research is based on positive and negative dimensions. Even though, the effect of positive and negative metrics of Twitter sentiment seems to be inconclusive (Bollen, Mao and Zeng's, 2011; Srivastava, 2012), this presents as an opportunity to tackle the issue. Furthermore, analysing news sentiment seems to be a slightly less attractive way of sentiment investigation. The previous studies specifically focus on negative sentiment (Tetlock, 2007) or news related to company earnings (Mian & Sankaraguruswamy, 2012). To my knowledge, there are not comprehensive empirical daily studies of news used as investor sentiment.

2.1.2 Search Engine Query

In recent years, Google has had available Google Trends tool, which allows the general public to collect research query data. For researchers and politicians, this tool has brought a new perspective on behaviour activities of the general public. This tool provides a volume ratio of searched queries (Vosen & Schmidt, 2011). The value of Google data is evident from the previous researches. For instance, Dugas, Jalalpour, Gel, Levin, Torcaso, Igusa, and Rothman, (2013) research suggests the possibility to predict the location of pandemic influenza before it starts. This was not possible with medical reports provided by a government agency because these reports are at least two weeks delayed. On the other hand, Google Trends website reports in real-time according to Choi and Varian (2012). They demonstrate it on visitor arrival statistic, where they compare official statistics with google search queries. Their method is able to predict the official statistics 6 weeks earlier. Because there is one-month lag preceding the official statistic and they can predict from the first two weeks of a month the rest of the month.

Vosen and Schmidt (2011) conclude that searched queries are a better predictor of consumption than a questionnaire-based indicator. Even though their research was conducted at the early stage of Google Trends tool development, when the tool did not support advance sampling, so they argue that the results might be more precise if the research was conducted recently.

Carrière-Swallow and Labbé (2013) forecast automobile purchase figures in Chile. Their rather satisfactory results demonstrate data reliability even in emerging markets, where the internet connection is not commonplace for masses.

Another important factor is the reliability of Google Trends data compared to questionnaires is Ripley's (2013) research quality of educational institutions. Where she establishes that the way how institutions administer the inquiry such as the ratio of students participating in the study, whether the students are cautious in filling up the questionnaires and if the institution sends the results on time are greater indicators of institution quality than data in the questionnaire itself. Apart from that, Stephens-Davidowitz and Pinker (2017) argue the primary shortage of a questionnaire that participants do not take it seriously, they might feel ashamed, or they might lie. In the end, the reason for bias is irrelevant, because, the results of questionnaires are biased one way or another. Furthermore, the research suggests that there is not an incentive to lie or falsify in the case of Google (Stephens-Davidowitz & Pinker, 2017). Also, people often share private information, which they would not tell even their best-friend or a doctor.

Previous research has shown the excessive potential of research query data, which scholars scratch the surface of in various fields. There is immense potential, because of the extensive large dataset with an enormously large group of participants, 3.5 billion searched queries daily, which can be obtained from Google Trends website (Marr, 2018, A). Therefore, the ambition is to uncover investor interest to trade a particular company.

Investor interest

There is not wide range of alternatives for investor interest proxies. The interest should reflect investor purchasing activity on stock markets. It means that the proxy must be correlated with stock volatility. Scholars have discovered few quantities correlated with stock volatility: major stock markets are linked (Ramchand & Susmel, 1998), bond market and S&P 500 are partly correlated (Fleming, Kirby & Ostdiek, 1998) also with real estate market (Mi, & Hodgson, 2018). However, according to Fleming, Kirby and Ostdiek (1998), these markets are in motion as a group, because of the same information, which is obtained by investors. This means that they react on the same subject in a similar manner, and the reaction of one market has limited prediction capability on another market.

Whereas, Bordino's et al. (2012) research indicate that online searched queries have prediction capability on market volatility. They use Yahoo! Search engine to compare searched queries with market movements. Their hypothesis is tested market-wide (NASDAQ-100 stocks) and on individual stocks. Market-wide results are far less significant (approximately 32 per cent) in comparison to an individual stock where correlation coefficients of 6 companies are higher than 70 per cent (Bordino et al., 2012). Furthermore, the coefficients are most robust the same day; this would suggest that investors evaluate markets and make immediate decisions. However, the (both positive and negative) one day lagged figures are likewise applicable, which imply that investors evaluate portfolio when the market moves as well as when new information emerges. This view is supported by Dimpfl and Jank (2016), who argue that the interaction is bi-directional: high (low) volume of search queries is followed by high (low) market volatility and vice versa.

In theory, today's search engine queries are able to forecast tomorrow's market volume volatility, in other words, they portray investor interest to trade stocks. However, they do not specify whether the price development is going to be in an upward or downward direction (Stephens-Davidowitz & Pabon, 2017). Therefore, combining search engine queries data with investor sentiment could provide insight into the future equity market development.

2.2 Theoretical Approach

It is rather complicated to predict the stock market due to its complexity. The stock market is influenced by factors such as political decisions, market conditions, annual reports and investors' expectations (Oh & Kim, 2002). The first assumption is that newspapers and Twitter reflect all of these initial inputs. In theory, investor sentiment should provide an approximate investors' attitude towards investing. In this scenario, it is not critical whether investors are rational thinking individuals who make a decision based on facts or they are characterised by herd behaviour, due to the fact, that newspapers and Twitter contain both information. Investors follow financial advisors on Twitter who gives equity market-related recommendations as well as CEOs who shares annual results and their visions. In a similar way, news addresses facts and recommendations investigated by journalists.

The second assumption is that if a stock price has been adjusted, investors will first search for more information in order to make a precise decision before purchasing new or reselling acquired stocks. Likewise, if an investor perceives new information which results in the shift of investor sentiment, the investor will research the topic and adjust a portfolio accordingly. Consequently, there will be a higher level of information interest preceding market volatility.

Research question: Can current investors sentiment combine with investor interest to trade forecasts the tomorrow's stock return?

The literature (Bordino et al. 2012; Wu & Brynjolfsson, 2015; Stephens-Davidowitz & Pabon, 2017) suggests a link between searching for a particular stock ticker and volatility of the stock. In other words, search engine query clustering should forecast market volatility.

Hypothesis 1: Higher interest in acquiring new information shows greater activity from investors to trade stocks.

It is essential to analyse which sentiment is more suitable to be proxy for investors sentiment whether Twitter sentiment or news sentiment. Bloomberg data shows mean values of news articles volume and Twitter post frequency where Twitter has a noticeably higher frequency of posts, and higher frequency data should be a more representative sample of average investor sentiment value. Even though the previous research is inconclusive, Srivastava (2012) results demonstrate strong cause-effect on stock return. In contrary, Bollen, Mao and Zeng (2011) do not discover it, and news sentiment has not been sufficiently analysed. Due to a lack of attention towards news sentiment and Twitter sentiment's higher data volume, Twitter sentiment might be a more appropriate proxy for investor sentiment than news sentiment.

Hypothesis 2: Twitter sentiment is more suitable as a proxy of aggregate confidence of investors towards equity markets than news sentiment.

Previous research in the field of behavioural economics, which combines two seemingly different fields economics and psychology, analyses decision making under risk. In 1979, Kahneman and Tversky (1979) introduced the prospect theory. This theory investigates decision making under risk. Their descriptive model suggests 2-2.5 more considerable aversion for losses than for gains (Kahneman, 2003). In other words, people indicate a higher sensitivity for losses. In terms of the equity market, this higher level of loss avoidance should result in surpassing the sensitivity of investors on negative aggregate confidence of investors towards stock markets. In addition, Browder (2015) has observed interconnectivity trough investor sentiment. It means that, if investors are experiencing negative returns on "market A", they tend to withdraw assets from "market B", although the markets are not connected. The negative sentiment seems to give preference for protecting capital above acquiring new assets. Lastly, Mian and Sankaraguruswamy (2012) confirm this theory in their investigation of stock market reaction on positive and negative company earnings results as well as Tetlock (2007) who finds a strong relationship between negative words in Wall Street Journal and market volatility.

Hypothesis 3: Negative sentiment is a stronger predictor of the stock market return then the positive sentiment.

Investor sentiment does not influence all stock prices equally. According to Baker and Wurgler (2007), stocks of high capitalization such as dividend-paying stocks, established companies with a transparent history and long-term profitability tend to be less influenced by investor sentiment. This means that the Apple stock return should be more difficult to forecast because it fulfils all the requirements. Even though, Amazon is not a young company without history. Amazon was unprofitable first 14 years since it was founded, and it does not pay dividends. Moreover, it had unprofitable quartiles between 2011 and 2015 (Griswold and Karaian, 2018). There should be a sensible difference between Apple and Amazon.

Hypothesis 4: The prediction accuracy is weaker on Apple stock return in comparison to Amazon stock return.

3 Data

The dataset used in this research has been collected from 2 sources: the Bloomberg Terminal and Google Trends website. The Bloomberg Terminal provides stock market-related data such as opening price, closing price, volume and sentiment analysis. Google Trends data shows the interest of investors in a current situation on a particular stock. The study is conducted through 3 years from January 2015 to December 2018. The dataset has been reduced by non-trading days. In total, there are 958 observations. However, only the Daily Return predicted variable has 958 observation. The rest of the variables was reduced by 1, because they are time-lagged variables.

3.1 Source Material

3.1.1 Bloomberg Data

As has been introduced in Previous Research section, this study is limited from choosing which data source to use. Due to the essence of this research, daily stock returns may be predicted with the sentiments which adjust daily, and as well it has the ability to predict. Consequently, this research will use two sentiments based on news articles and Twitter micro-blog posts computed by Bloomberg.

Bloomberg is a worldwide recognized source of market-related data for investment banks, hedge fund managers and other finance professionals. The core of their business strategy is based on being reliable data provider with 325,000 subscribers (Bloomberg, 2019). Moreover, they have developed a sentiment analysis tool. The tool reads trough thousands of news and Twitter posts and generates quantitative numerical output in real-time (Bloomberg, 2016). Bloomberg does not disclose how exactly this tool works, apart from using the machine learning algorithm. On the other hand, the tool has been developed by a team of PhDs (Bloomberg, 2014). This fact and their reputation are strong predictors of creating one of the best sentiment sources.

Moreover, there are not many alternatives to the Bloomberg Terminal. For instance, OpinionFinder used by Bollen, Mao, and Zeng (2011) in a similar study is a freeware, which can detect subjectivity from the text (Wilson, Hoffmann, Somasundaran, Kessler, Wiebe, Choi, & Patwardhan, 2005) similarly as Bloomberg. However, this method would be time-consuming and presumably without acquiring more reliable data. Since their reputation as a world leader in the industry is not in stake and companies such as Bloomberg has generous budged for the data research. Another way is to develop a new algorithm which would have comparable disadvantages as OpinionFinder. Lastly, the Bloomberg Terminal is a widely accepted tool for market-related data, trusted even by Goldman Sachs bank. Therefore, Bloomberg's sentiment analysis should be one of the best publicly accessible sources of investor sentiment.

Lastly, Bloomberg provides Twitter sentiment and news sentiment data. Owing to the fact, both sentiments are analysed in this study and they can be from one source is beneficial. Data from one source is more comparable and reduces bias caused by different approaches of collecting and computing the final values.

3.1.2 Google Data

As was described in the previous chapter, there are not abundance alternatives for investor interest proxy. Even though, scholars have found link between stock volatility of major markets among themselves (Ramchand & Susmel, 1998), bond market and S&P 500 (Fleming, Kirby & Ostdiek, 1998) and real estate market (Mi, & Hodgson, 2018). Those are perceptibly co-movements, however, some theorists (Bordino's et al. 2012; Stephens-Davidowitz, & Pabon, 2017) argue that search engine queries are capable of forecasting market volatility. It appears that there are not any sufficient alternatives. With respect to it, the analysed data could be from Google, Yahoo!, or Bing engines. A representative sample requires participants, Google search engine's market share is nearly 90 per cent, in comparison to 6 per cent Bing and 4.5 per cent Yahoo! in the United States of America (Statcounter, 2019). Preferable, Google data seems to be a surpassingly ingenious option.

Furthermore, Preis, Moat, and Stanley (2013) conclude that Google Trends data are capable of predicting future market movements. Their research as well suggests that data collected exclusively from the United States of America better reflect the situation on the American equity markets than the world-wide data. Hence, this research will use Google Trends data collected exclusively from the United States of America.

The Google Trends tool seems to be the most suitable for this research, in spite of two essential drawbacks. First, it does not provide the total numbers. Second, it is complicated to research period exceeding 7 months. Google Trends provides only scaled proportion in a range 0 to 100, which requires to round the figures. This creates rounding bias. Even though, this research might be conducted with total figures. More importantly, Google Trends tool does not support the collection of data on a daily basis for a longer period of time than 7 months. If the researched query period is longer than 7 months Google Trends automatically shows weekly data instead of daily data. Due to the first issue, it is not possible to download a 7 months period and add a subsequent 7 months period. That would result in rescaling of the final 3 years dataset. To overcome this issue the 7 months periods were back spliced into one final dataset with one month overlap period. This method partly removes scaling bias.

4 Methods

This paper combines two proxies: investor sentiment and investor interest to predict stocks of Apple and Amazon. Investor interest gives the impression of being linear with market volatility with regards to Bordino's et al. (2012) research and the replication of their research in the next chapter (please see the cross-correlations tables and figures in 5.1.1 and 5.1.2 sections). On the contrary, investor sentiment has a non-linear relationship with a stock market (Bollen, Mao & Zeng, 2011). Therefore, Apple and Amazon stocks will be forecasted with an Artificial Neural Network (ANN) model, which has, according to Guresen, Kayakutlu and Daim (2011), the most reliable and realistic predictions in predicting time series.

4.1 The Approach

An ANN is a computer algorithm inspired by a human brain, it has self-learning capability. For instance, it can learn to distinguish the difference between pictures of dogs and cats. It must be provided a large data set of pictures tagged "dog" or "cat" accordingly, in order to teach the computer which elements of a picture are important for a determination whether there is a dog or a cat (Marr, 2018, B). In contrary, a classic algorithm, where the programmer has to discover a pattern and then implement it as a code. Interestingly, ANN algorithms can become surpassingly advance that even the original author does not understand the logic behind their decisions (Beardsworth & Kumar, 2019). This capability is incredibly favourable for excessively comprehensive issues such as forecasting movement of stocks. Even though this research is not exceeding a 3 years period (approximately 960 observations) and it has only 5 input variables, it would be genuinely complicated to estimate the wages of individual variables.

4.2 Model Specification

Generally, the ANN model must be adjusted accordingly to its purpose. Figure 1 shows a simplified diagram of the ANN model applied in this research, the real model would be difficult to present on the paper due to its large size. The input layer contains 5 inputs: positive and negative Twitter sentiment, positive and negative news sentiment and investor interest. This deep neural network has 2 hidden layers as applied in similar research by Wanjawa and Muchemi (2014), which consist of 32 neurons in the case of Apple and 16 in case of Amazon, instead of 4 as displayed in Figure 1. The output layer compares the learnt results with the real daily market return values. Lastly, the learning mechanism consists of 500 epochs, one epoch is one-time used dataset. Each neuron adjusts the weight of the input variables. Moreover, it has been adopted a classic split 80/20, where 80 per cent of data is a training dataset and 20 per cent to test the learnt algorithm, which is an accepted rule by the research community and applied in similar studies, for instance, Wanjawa and Muchemi (2014).

This means that the inputs are fed into the input layer 500 times. Neurons' weight is calculated by neurons in hidden layers with an activation function (Missing Link, 2018). This ANN model uses Rectified Linear Unit (ReLU) activation function. ReLu is popular activation function in recent years, due to several advantages over its predecessors such as sigmoid and hyperbolic tangent functions. Both predecessors have a losing gradient issue. Therefore, they are not applicable to deep learning (Brownlee, 2019). In addition to it, more importantly, ReLU superiority is in computational simplicity (Brownlee, 2019).

Lastly, an output of each neuron is calculated through the equation below, where x_1 is value of output of the neuron in the previous layer and "w_1" is its computed weight, "x_2" and "w_2" are values from a distinct neuron in the previous layer, this continues until all output values from the previous layer are taken into the equation. For the models used in this work, it would be: "x_n" equals 5 for the first hidden layer, and the second hidden layer and output layer receive signals from 16 neurons in the case of Apple and from 32 in the case of Amazon.

$$output = ReLU(x_1 w_1 + x_2 w_2 + \cdots + x_n w_n + bias)$$

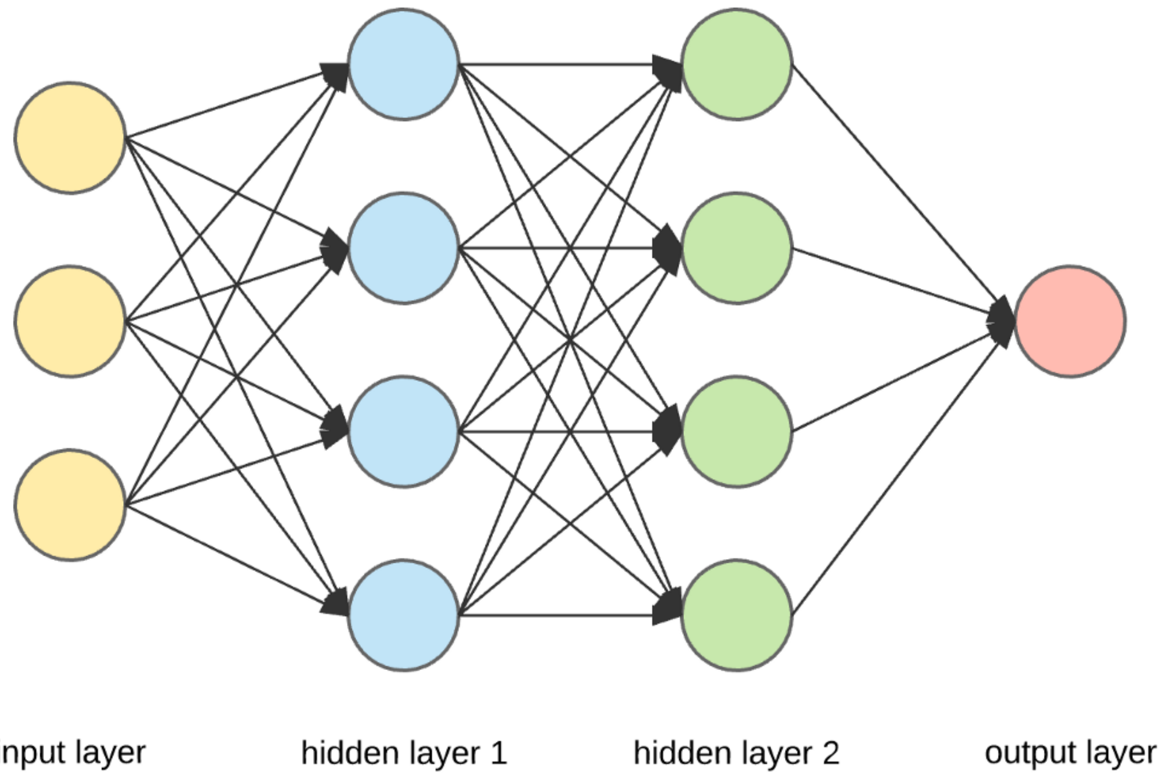


Figure 1 Artificial Neural Network Model (Dertat, 2017)

4.2.1 Predicted Variable

The predicted variable measures the daily market return. There are two methods which measure daily stock change: open-to-close and close-to-close daily change, which can be as a percentage value or a total number. This research uses PCH, which is daily stock return measured in percentage. PCH is an open-to-close growth metric calculated according to the equation below. "C" represents closing price and "O" is open price. The reason why to use percentage change is argued in the next chapter, Extension part.

$$PCH = \frac{C - O}{O}$$

According to Li, Xie, Chen, Wang, and Deng (2014), there are two reasons to apply the open-to-close metric instead of close-to-close. Firstly, weekend-gap is associated with higher risk. The risk may be overcome by withdrawing investment before the gap and reinvest afterwards. Thus, investors are more active at the begging and at the end of weekdays. This activity does not have an impact on the open-to-close metric. Secondly, seasonality of interday traders. Equity markets have busy seasons and inactive seasons such as summer holiday. These interday traders might bring inexplicable patterns into this research (Li et al. 2014).

4.2.2 Inputs

There are 2 types of inputs: investor sentiment and investor interest, however, investor sentiment is constructed from 4 sentiments variables: positive Twitter, negative Twitter, positive news and negative news. Moreover, all inputs are time-lagged by -1 (t_{-1}), as the aspiration is to predict stock movement one day in advance. All negative inputs (Twitter and news negative sentiment) were transferred into positive variables due to ReLU function inability to deal with negative or zero numbers (Missing Link, 2018).

Investor Sentiment

As is disclosed in the introduction, there are several investor sentiment proxies, however this research requires daily basis change. Therefore, this paper analyses secondary data collected by Bloomberg. The main disadvantage is not a publicly available algorithm, which was applied to collect the data. Their algorithm might bias this research, however, since it was developed by a team of PhDs (Bloomberg, 2014) it should be reliable. The bias might be in adjusting the algorithm during the researched period, which would result in an inaccurate learning or testing process of the ANN model. In theory, this issue could be overcome by developing a new algorithm or finding open-source software. In practice, the assumption is that Bloomberg's software is excessively advance (precise) in comparison with an open-source solution.

Investor Interest

Google Trends tool has been used to collect investor interest data. Naturally, all the people searching the name of a company are not investors, especially in the case of Amazon, where millions of customers aim to purchase products, not shares. Another option could be researching queries such as "Amazon news", "Apple stock" and "Amazon return", however, Bordino et al.

(2012) compare company names and stock tickers, and the stock tickers' cross-correlations have significantly stronger coefficients. Given their favourable results, a stock ticker seems to be a reasonable query to analyse. Moreover, the appropriate word should be timeless according to Dzielinski, (2012), something that is not seasonal. Stock ticker seems to have this capability. For instance, the phrase "Apple news" might have been seasonal when Apple lunch new product. Because it is not evident whether a customer researches new products or an investor seeks news related to Apple company. This phrase would possibly forecast product sales better than stock volatility.

5 Empirical Analysis

This chapter introduces relationships between variables and summary of their values. Furthermore, it replicates Bordino's et al. (2012) research in order to demonstrate the suitability of Google query data as a proxy for investor sentiment. Finally, is shows this research results and provides discussion.

5.1 Verification

The tables 1-2 give the summary statistics of Apple and Amazon's variables. It is noticeable that the average positive and negative values of Twitter are more balanced, whereas news has the tendency to be more positive. In figures, the probability of Apple news to be positive is 2.6 more than negative news in case of Amazon this probability is 4.3. Appendix A presents the interaction between Apple individual variables and Appendix B display it between Amazon variables. The first row is noteworthy because it shows the interaction of a daily stock return with inputs. Initially, any of the input variables do not indicate a strong relationship with a daily stock return.

Table 1 Summary Statistics of Apple Twitter Variables

Apple Variables	**Obs**	**Mean**	**Std. Dev.**	**Min**	**Max**
Predicted Variable					
Daily Return	958	4.28E-05	0.012524	-0.06633	0.086961
Inputs Variables t_{-1}					
Tweets Positive Count	957	399.2079	668.4419	10	7780
Tweets Negative Count	957	420.236	771.6369	9	12621
News Positive Count	957	127.0052	127.7585	9	1184
News Negative Count	957	48.5298	49.63261	1	418
Google Query Volume	957	30.7372	13.82201	14	115

Table 2 Summary Statistics of Amazon Twitter Variables

Amazon Variables	**Obs**	**Mean**	**Std. Dev.**	**Min**	**Max**
Predicted **Variable**					
Daily Return	958	-0.00028	0.01506	-0.08561	0.07452
Inputs Variables t_{-1}					
Tweets Positive Count	957	192.2685	379.1989	10	5637
Tweets Negative Count	957	164.682	334.0946	3	4899
News Positive Count	957	63.88192	61.38059	3	531
News Negative Count	957	14.7994	17.82761	1	188
Google Query Volume	957	17.82106	14.88508	1	125

5.1.1 Replication

This study uses Google query data as a proxy for investor interest to trade a certain stock. Studies such as Wu and Brynjolfsson (2015) and Bordino et al. (2012) find a strong relationship between searched queries and market volatility. However, these figures vary from a company to a company, which could have had a significant impact on the result of this study. In this section, the exact level of correlation for Apple and Amazon, it is identified. The replication methodology is inspired by Bordino's et al. (2012) research.

From examining the recent studies, Google Trends data seems to be a decent source of information on human behaviour (Stephens-Davidowitz, & Pabon, 2017). Each time an investor researches new potential opportunity or evaluates the current portfolio, it leaves an online trace. The trace reveals the possible intention of trading, which might be used to forecast the stock market volatility (Wu & Brynjolfsson, 2015). Bordino et al. (2012) conducted a research where they found a correlation between a stock ticker and volatility of the stock. Their research suggests that a stock ticker average (NASDAQ-100 stocks) cross-correlation coefficient is 0.3150 in comparison to a company name where the value is 0.1196 (Bordino et al. 2012).

Furthermore, here is an enormous deviation between average coefficients and a single company coefficient. For example, Adobe Inc 0.83 coefficient is much greater than the average value. For this reason, it is necessary to replicate their research for Apple and Amazon, in order to determine their relationship between a stock ticker query and a company volume volatility. Lastly, Bordino's et al. (2012) research was conducted with the Yahoo! search engine on contrary this research will use Google data. This fact might influence the results.

Appendix C shows the relationship between the searched stock ticker query and daily stock volume, which is rather illegible. Because Appendix C displays 3 years of overall results. On the other hand, Figures 2 (Apple) and 3 (Amazon) present the first 100 days, where the correlation is apparent.

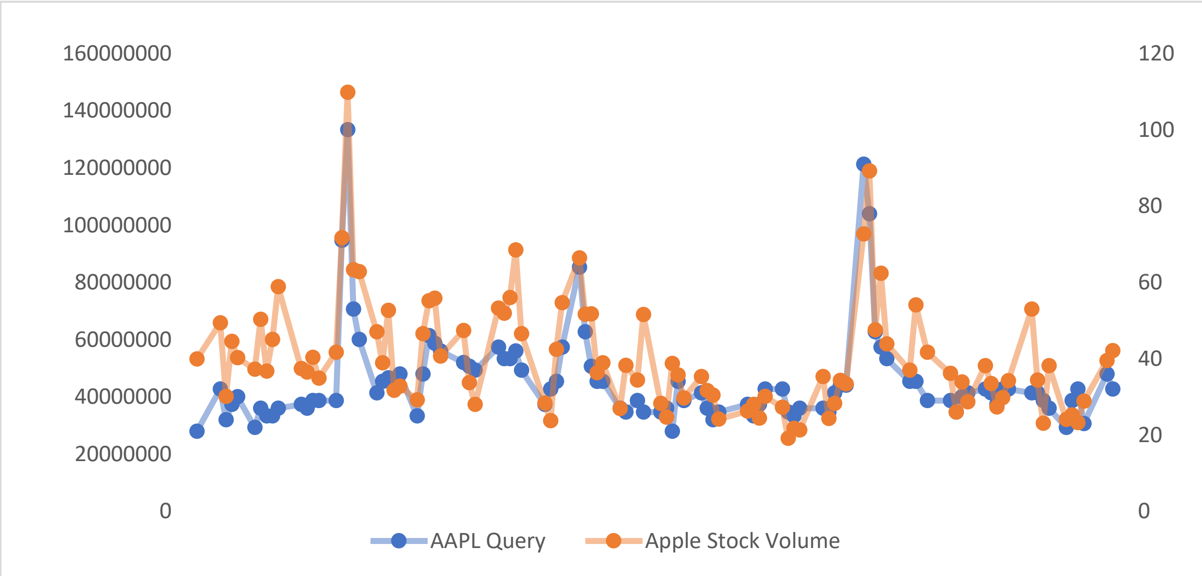


Figure 2 Correlation of Apple Stock Volume and AAPL Query (100 days) , Data Source: Google Trends and The Bloomberg Terminal, Note: Author's Calculation

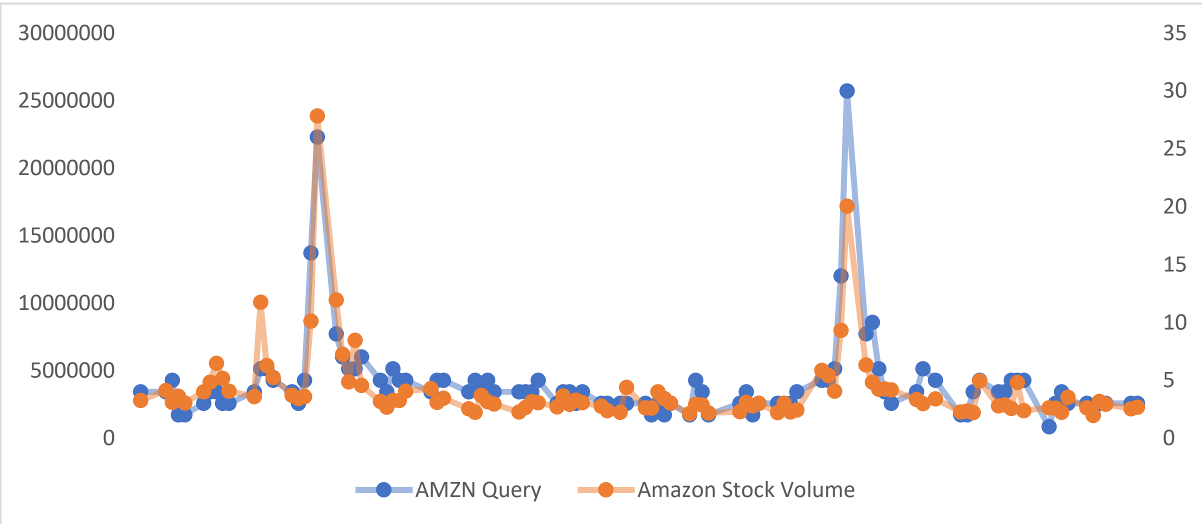


Figure 3 Correlation of Amazon Stock Volume and AMZN Query (100 days) , Data Source: Google Trends and The Bloomberg Terminal, Note: Author's Calculation

As in the original research by Bordino et al. (2012), it is used time-lagged and time-forwarded Pearson correlation, where the sample Pearson correlation coefficient r(δ) is defined as:

$$r(\delta) = \frac{\sum_{i=1}^{n}(Q_i - \bar{Q})(T_{i+\delta} - \bar{T})}{\sqrt{\sum_{i=1}^{n}(Q_i - \bar{Q})^2}\sqrt{\sum_{i=1}^{n}(T_{i+\delta} - \bar{T})^2}}$$

"n" is a sample size of approximately 950 observation. "Q_i" represents the volume of stock ticker queries "AAPL" and "AMZN". "$T_{i+\delta}$" is Apple and Amazon daily trading volume where "δ" is a day time-lag with values from -5 to 5. The time shift demonstrates how relevant are data from previous days on stock volume.

Table 3 Cross-Correlation of Stock Volume and Searched Queries

Day (δ)	δ=-5	δ=-4	δ=-3	δ=-2	δ=-1	δ=0	δ=1	δ=2	δ=3	δ=4	δ=5
Company											
AAPL	0.1242	0.1515	0.2030	0.3038	0.4738	0.6815	0.5196	0.2810	0.2062	0.1647	0.1525
AMZN	0.2566	0.2801	0.3035	0.3414	0.4643	0.6558	0.5381	0.3821	0.3255	0.2833	0.2643

Cross-correlation in Table 3 shows a similar pattern as in the original study by Bordino et al. (2012). For comparison reason, their results are in Appendix D. The coefficients gradually grow till the peak at lag δ=0, then decrease. The peak values are 0.6815 (Apple) and 0.6558 Amazon, which is a moderate positive correlation between searched queries and stock volume for both companies. Bordino et al. analyse 87 clean stocks, and both Apple and Amazon would be among the top ten most cross-correlated companies. Apple would be on the 7th place with Netflix and Amazon on 10th would share its position with F5 Networks.

The correlation results might be driven by outliers. Therefore, the results are tested against the replication of the research without outliers. For example, the coefficient of Apple at lag δ=0 is 0.6546 without outliers and the value with outliers is 0.6815. This 0.0269 insignificant difference suggests that the correlation is not driven by peak events such as the big announcements of new products or financial reports, according to Bordino et al. (2012).

5.1.2 Extension

In theory, stock volume volatility should lead to stock price volatility. Because in economic theory, demand influences supply and vice versa. According to Campbell Grossman and Wang (1993) there are two reasons for a stock price change: new information has arisen, which has adjusted investor sentiment, or it might be because of traders who become sensitive for loss aversion. Even though previous researchers have acknowledged the relationship (Campbell Grossman and Wang (1993); Cont (2001); Liu, Gopikrishnan & Stanley (1999)), it is necessary to understand whether the relationship for Apple and Amazon is robust.

Since the investigation regarding stock price and daily stock volume is very similar to Google search queries and stock volume, the analysis of this relationship proceeds in a similar manner as in the previous section. Daily stock volume volatility is compared with daily absolute values of stock price change. Stock price change had to be converted into absolute values because daily stock volume values can solely be positively associated. CHAbs (daily stock price change in absolute values) is a formula used to calculation of daily stock price change where C represents closing price and O is opening price.

$$CHAbs = |C - O|$$

Figure 4 and 5 show rather suspicious output where is a disparity between the proportion of stock volume and daily price change. Especially Amazon demonstrates high stock volume volatility and low daily price change 3 years ago. The picture is inverse in the recent year. These figures suggest an inaccurate measure of daily price change. Because the opening price of Amazon was $312.58 at the begging of 2015 and it was $1510.8 at the end of 2018. Due to this astounding growth of the company in 3 years, the daily price change must reflect the percentage growth. To create PCHAbs (absolute stock price change in %) require only modest adjustment in the CHAbs equation.

$$PCHAbs = \left|\frac{C - O}{O}\right|$$

Figure 6 and 7 demonstrate a correlation between PCHAbs and stock volume volatility. Amazon shows tremendous improvement (see Figure 5 and 7). The values seem to be equally spread with PCHAbs equation.

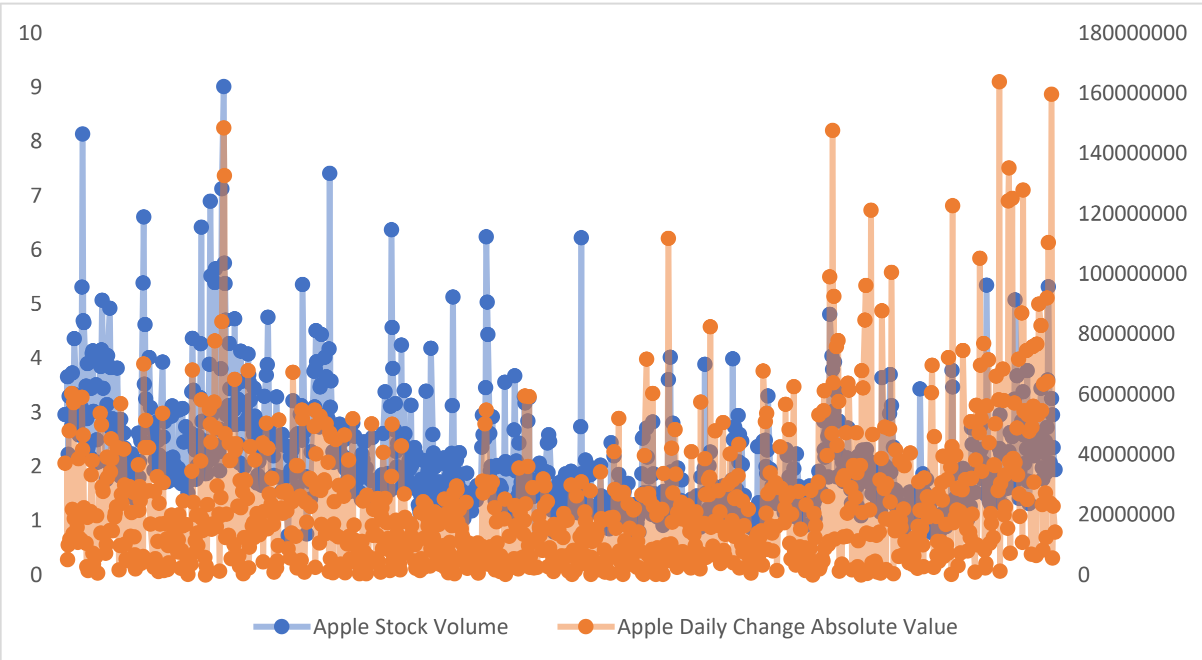


Figure 4 Correlation of Apple Stock Volume and Apple Daily Price Change (3 years), Data Source: Google Trends and The Bloomberg Terminal, Note: Author's Calculation

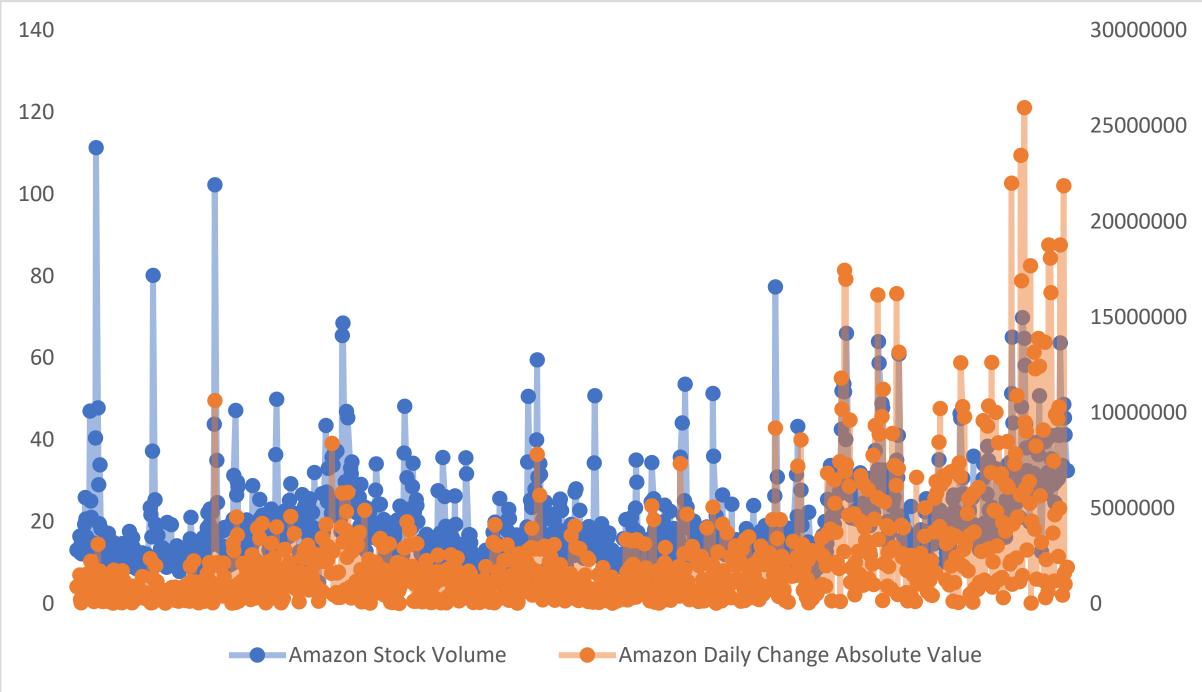


Figure 5 Correlation of Amazon Stock Volume and Amazon Daily Price Change (3 years), Data Source: Google Trends and The Bloomberg Terminal, Note: Author's Calculation

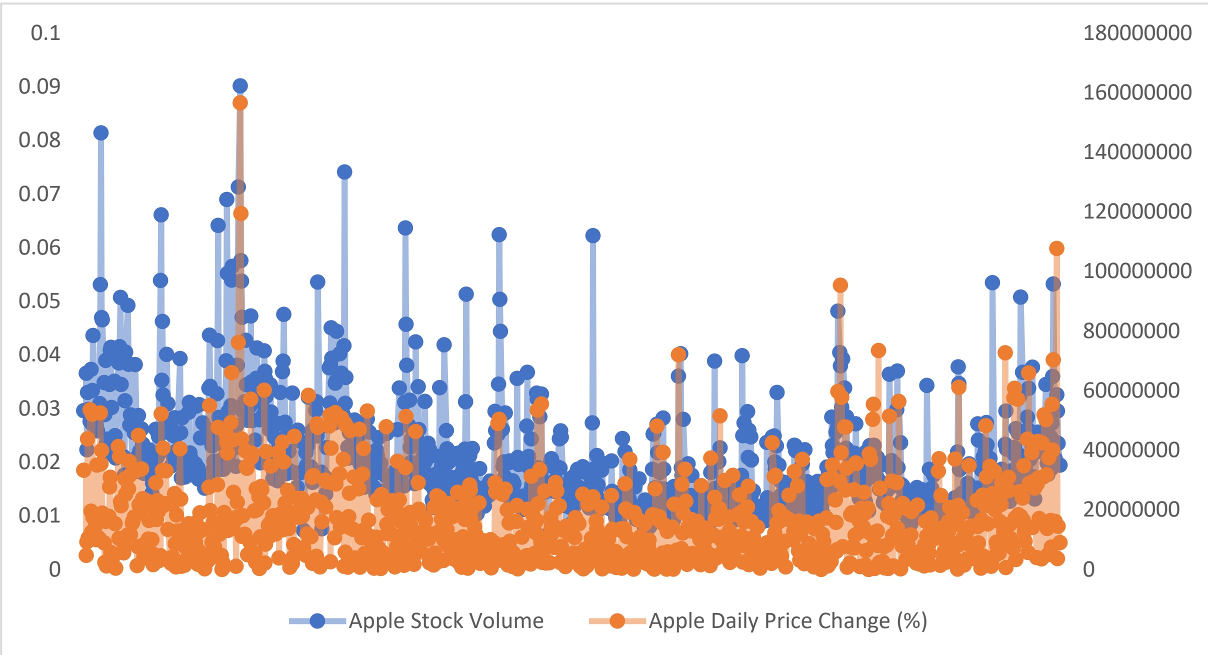


Figure 6 Correlation of Apple Stock Volume and Apple Daily Price Change in % (3 years), Data Source: Google Trends and The Bloomberg Terminal, Note: Author's Calculation

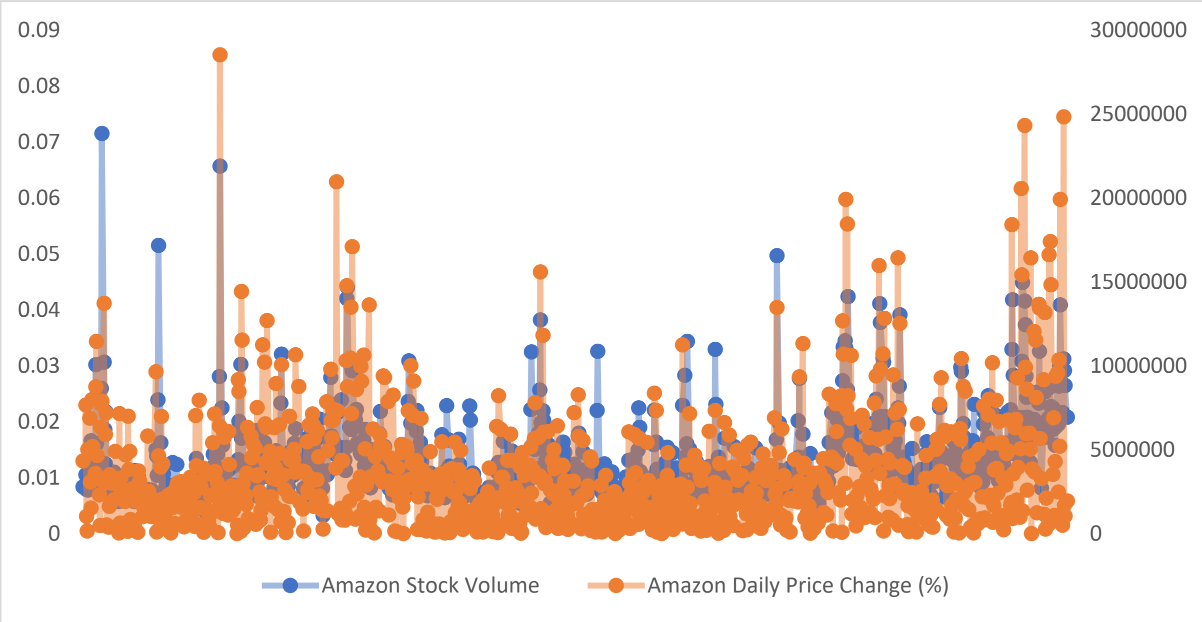


Figure 5 Correlation of Amazon Stock Volume and Amazon Daily Price Change in % (3 years), Data Source: Google Trends and The Bloomberg Terminal, Note: Author's Calculation

Table 4 shows the results of cross-correlation between stock volume volatility and stock price change in %. Table 4 is comparable with Table 3 in regard to the pattern, and even though the cross-correlation is slightly less substantial. This suggests that there is a stronger relationship between Google searched queries and stock volume volatility than stock volume volatility and stock price change in %.

Table 4 Cross-Correlation of Stock Price Change in % and Stock Volume Volatility

Day (δ)	**δ=-5**	**δ=-4**	**δ=-3**	**δ=-2**	**δ=-1**	**δ=0**	**δ=1**	**δ=2**	**δ=3**	**δ=4**	**δ=5**
Company											
AAPL	0.1404	0.1468	0.1911	0.2742	0.3469	0.5663	0.3770	0.2736	0.2221	0.1972	0.1520
AMZN	0.1995	0.2493	0.2685	0.3064	0.3427	0.5871	0.4402	0.3052	0.2651	0.2393	0.2063

Table 5 is a control table where the absolute price changes are cross-correlated with stock volume volatility. The cross-correlation is less substantial in comparison with Table 4 where the % approach is demonstrated. In light of the evidence, stock % growth price is more sufficient measurement of stock price change.

Table 5 Cross-Correlation of Stock Price Change and Stock Volume Volatility

Day (δ)	**δ=-5**	**δ=-4**	**δ=-3**	**δ=-2**	**δ=-1**	**δ=0**	**δ=1**	**δ=2**	**δ=3**	**δ=4**	**δ=5**
Company											
AAPL	0.0523	0.0647	0.1102	0.1773	0.2334	0.4287	0.2657	0.1753	0.1328	0.1215	0.0735
AMZN	0.2540	0.2985	0.3075	0.3509	0.3612	0.5470	0.4577	0.3591	0.3341	0.3067	0.2738

This replication has confirmed Bordino's et al. (2012) research. The results from the replication part – between searched stock ticker query and daily stock volume – are comparable with Bordino's et al. (2012). Apple and Amazon do not show as strong relationship as Adobe (compared with Appendix D). However, their relationships are robust. The extension part has uncovered the link between stock volume volatility and stock daily return. Therefore, search Google query data seems to be a suitable proxy for investor interest. Lastly, this part has brought to the light improved measurement of daily stock return. Consequently, the daily stock return measured in percentage is applied for the primary research of this thesis.

It is noteworthy, to point out how comparable Apple and Amazon figures are in this section. The final ANN model will use the δ=-1 variable as an input, where the correlation coefficients are 0.4738 (Apple) and 0.4643 (Amazon) for search engine queries and stock volatility. In the case of stock volatility and stock price change in % the coefficients are truly similar to 0.3469 (Apple) and 0.3427 (Amazon). These equal values establish a feasible foundation for the actual research, where it will be possible to observe interaction among variables without bias caused by a stronger investor interest input. This part suggests that investor interest input variables will have a comparable influence on Apple and Amazon return prediction.

5.2 Results

Firstly, this section demonstrates the learning and testing process of the ANN model. Secondly, it presents the relevance of individual inputs. Lastly, it compares this research's results with the literature.

The figures 8 and 9 display correlation between real stock return and computed value during the learning period. These figures are zoomed in due to readability reasons. The whole learning period (approximately 760 observations) can be found in Appendix E. The correlation coefficients are 0.40 Apple and 0.52 Amazon. This means that the ANN model has been able to identify a pattern, which hardly ever do not appraise a correct direction, however, it has repeatable obstacle to estimating the precise size of daily gains or losses (see the figures 8 and

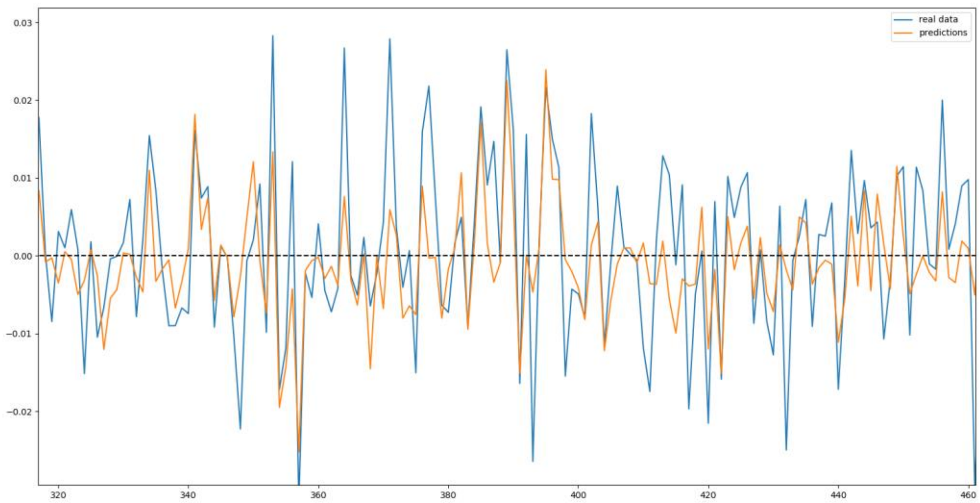


Figure 8 Apple Daily Return, ANN Model Training (Zoomed), Real Values in Blue and Predicted Values in Orange, Data Source: Google Trends and The Bloomberg Terminal, Note: Author's Calculation

9).

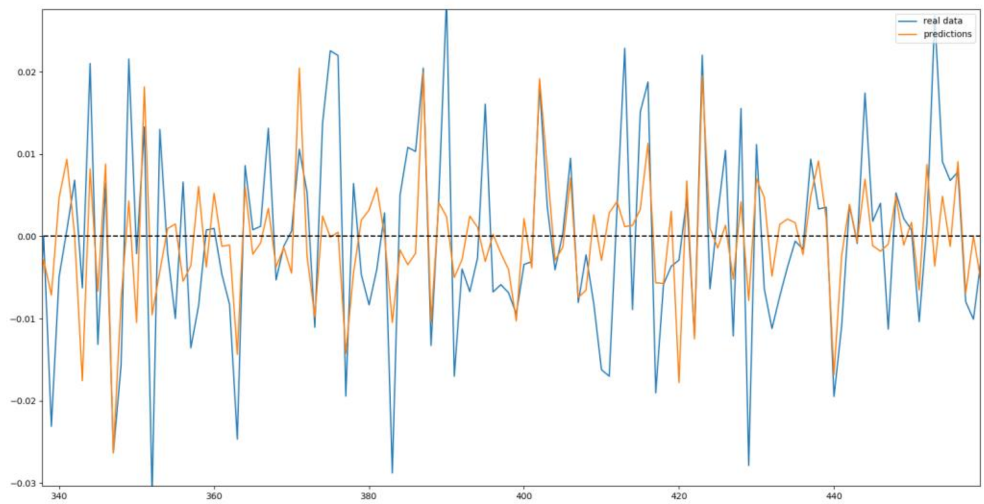


Figure 9 Amazon Daily Return, ANN Model Training (Zoomed), Real Values in Blue and Predicted Values in Orange, Data Source: Google Trends and The Bloomberg Terminal, Note: Author's Calculation

Then approximately 180 observation are tested by the learnt algorithm of the ANN model, the figures 10 (Apple) and 11 (Amazon) present comparison of real daily return with forecasted values (Zoomed in figures can be found in Appendix F). From these diagrams, it is presumable that such a complex issue as the daily market return cannot be forecasted only with investor sentiment and investor interest. Represented in numbers, the correlation coefficients are 0.02 (Apple) and 0.13 (Amazon). These results suggest omitted variable bias or as some theorists argue that the equity market is unpredictable. The histograms comparing prediction errors during learning and testing stages (see Appendix G) and the degradation is critical.

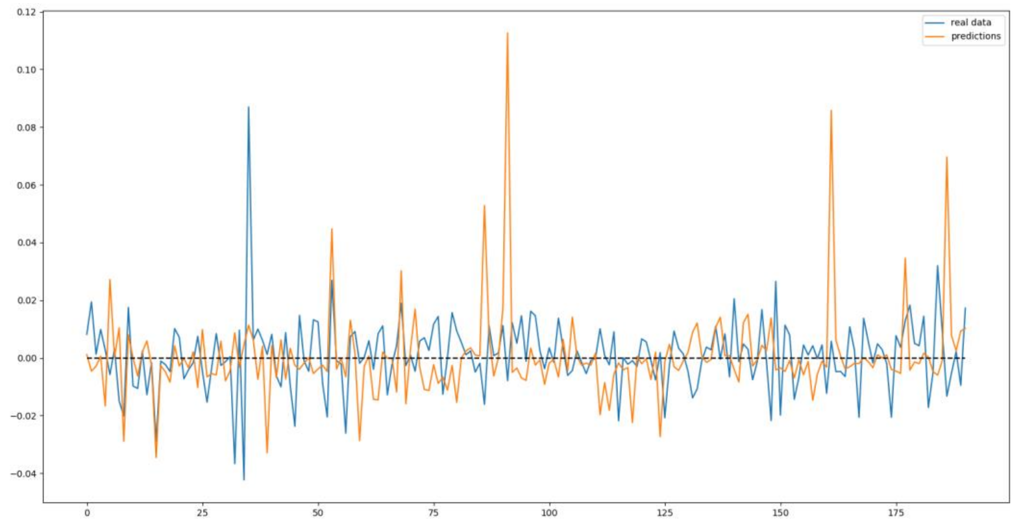


Figure 10 Apple Daily Return, Predicted Values by the ANN Model, Real Values in Blue and Predicted Values in Orange, Data Source: Google Trends and The Bloomberg Terminal, Note: Author's Calculation

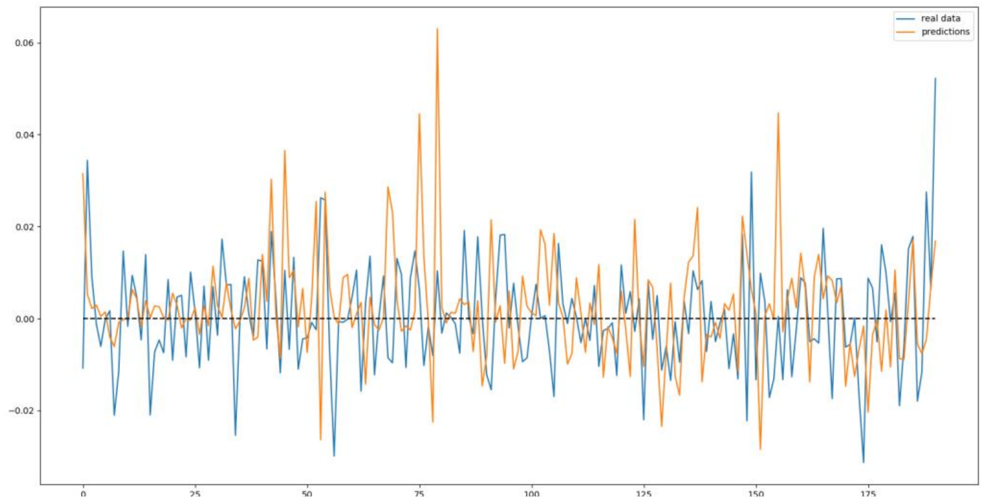


Figure 11 Amazon Daily Return, Predicted Values by the ANN Model, Real Values in Blue and Predicted Values in Orange, Data Source: Google Trends and The Bloomberg Terminal, Note: Author's Calculation

Possibly the most serious drawback of an ANN analysis is the interpretation of how relevant are its inputs. Notably, in comparison to Ordinary Least Squares regression method, where relevance of each independent variable can be measured by its weight. Since an ANN model is a non-linear, the weight of input is connected to hidden layers (Ryan, 2018). These weights can be obtained in the form of an excessively long equation, the equation length is related to the number of layers and neurons used in the ANN. Due to its complexity and length, this equation is not clearly arranged. Therefore, the ANN components importance can be measured indirectly via random value errors method.

This method feeds one input of the ANN model with random values then calculate mean squared error. This procedure repeats until all inputs mean squared errors are calculated. More massive error represents greater weight of the component. Table 6 displays the mean squared errors of investors sentiment and investor interest. Both Apple and Amazon suggest a similar pattern of individual component importance. In addition, positive and negative errors of news sentiment indicate equal importance. This not apply for Twitter, where negative sentiment is 3 times stronger predictor than positive sentiment. This means that the least essential input is positive sentiment obtained from Twitter. Overall, news sentiment and investor interest seem to be approximately twice stronger predictor in case of Amazon compared to Apple.

Table 6 Random Value Errors, (Mean Squared Errors and Mean Absolute Errors)

	Apple	**Amazon**
Inputs (components)	**Mean Squared Error**	**Mean Squared Error**
Twitter Positive Sentiment	1.4641e-04	1.8546e-04
Twitter Negative Sentiment	4.6534e-04	4.5899e-04
News Positive Sentiment	3.1893e-04	6.1054e-04
News Negative Sentiment	3.3386e-04	6.5759e-04
Investor Interest	2.8061e-04	5.3640e-04

5.3 Discussion

It should be noted that findings in Table 6 are based on training data. Inputs of both companies demonstrate a similar pattern of inputs relevance. This fact suggests two possible outcomes: the computing task was excessively complex, and the model suffer from omitted value bias or methodology bias which influence both results similarly. In the case of an excessively complex task, the findings of this study are presumably applicable in general. In the case of methodology bias, the influence of inputs on stock return is possibly accidental.

To tackle this issue, the ANN model reliability has been tested on stock closing price (forecasted variable) with the same inputs, and on the top of that stock closing price lagged t_{-1}. The results of this trivial task are extraordinarily strong, the correlation coefficients of predicted value and actual value are 0.98 in the learning phase and 0.97 in the testing phase. This additional investigation illustrates that the ANN model used in this research has the capability to learn and predict. Therefore, it is presumably the first outcome, that it is excessively complex to calculate the exact daily stock return, as some authors suggest impossible. Consequently, the results in Table 6 seem to be valid.

Furthermore, these results are comparable with literature. As Baker and Wurgler (2007) argue investor sentiment influence stocks, which value is difficult to estimate. Even though Amazon was profitable during the period of this research, and it had broken even at the begging of the research period. Thus the results are in agreement with Baker and Wurgler (2007) because news investor sentiment relationship with forecasted value is twice stronger for Amazon. Besides, Amazon's forecasted returns are 6.5 times more correlated with the actual returns than Apple's returns. It might be intriguing to test this hypothesis on a young company with no profit or history. However the Bloomberg Terminal does not provide sentiment data on insignificant companies.

In 1979, Kahneman and Tversky defined prospect theory, which analyses decision making under stress. Thereafter Kahneman (2003) analyses it and his descriptive model suggests 2-2.5 larger aversion for loss than for gains. The values in Table 6 display a similar pattern, where loss aversion of Twitter sentiment analysis shows 2.47 (Amazon) and 3.18 (Apple) times stronger aversion for losses than for gains. The loss aversion is inconsiderable in the case of news sentiment 1.08 (Amazon) and 1.05 (Apple). The link might be in the essence of those vehicles, news is obligated to report independently without any hidden interest, whereas tweets share personal opinions. This might mean that rational people read the news then decide accordingly with facts, and sentimental people browse through tweets and make decision-based on their current state of mind. However, this dissimilarity of Twitter and news sentiment requires further research.

In theory, investor interest should have analogous effect on Apple and Amazon. Due to their almost identical results in Replication part. As Table 3 shows the correlation between Google search query and stock volatility, the coefficients (δ=-1) are 0.4738 Apple and 0.4643 Amazon. Table 4 displays correlation (δ=-1) of stock volume volatility and stock return (in %) 0.3469 Apple and 0.3427 Amazon. However, Table 6 shows the different influence of investor interest in the case of Apple (2.8061e-04) and Amazon (5.3640e-04). This might be due to the fact that investor interest component is not able to forecast whether the stock has downward or upward direction although it can predict volatility. This might mean that investor sentiment level of significance depends on investor sentiment accuracy. In this relationship, investor sentiment determines the direction and investor interest adds the volume.

6 Conclusion

The effects of Artificial Intelligence on a stock market will certainly become an interest of studies from the academic sphere and financial institutions. Although, the ANN model has not been able to predict daily stock returns with investor sentiment and investor interest inputs. The additional research suggests the correct methodology with omitted variable bias. The issue seems to be excessively complex to be using only 5 input variables.

Even though the research aim has not been fully accomplished, due to the ANN models inability to precisely forecast daily stock returns, the models are capable to predict the positive or negative direction of the stock returns fairly accurately. However, the models lack capability to predict the precise value of the daily return. Therefore, the hypotheses shall be answered with some limitation.

Hypothesis 1: Higher interest in acquiring new information shows greater activity from investors to trade stocks. A large part of this research focuses on clarifing the issue. The cross-correlation in Tables 3 and 4 show a positive relationship between investors searching for company-related information, stock volatility and stock return respectively. The relationship is the strongest on the current day, however it suggests that investors analysis stock before purchase (see δ=-1 values) and evaluate afterwards (see δ=1 values). The 1st hypothesis is accepted.

Hypothesis 2: Twitter sentiment is more suitable as a proxy of aggregate confidence of investors towards equity markets than news sentiment. Table 6 provides a summary of mean squared error of inputs. The table indicates that Twitter sentiment can predict a downward tendency, however, Twitter positive sentiment is the least influential input. On the other hand, news sentiment, in general, is more influential on daily stock return. Notably, in the case of Amazon, news sentiment is a significant predictor of future market returns. This might imply that Twitter is largely occupied with less analytical or passionate investors and the news is used by more conservative and rational thinking investors. The 2nd hypothesis is rejected.

Hypothesis 3: Negative sentiment is a stronger predictor of the stock market return than the positive sentiment. There is a significant difference between Twitter’s positive and negative sentiment, where negative sentiment is approximately 3 times more influential on stock return than positive sentiment. In addition, it is also marginally stronger in the case of news sentiment. The 3rd hypothesis is accepted.

Hypothesis 4: The prediction accuracy is weaker on Apple stock returns in comparison to Amazon stock returns. The figures 10 (Apple) and 11 (Amazon) present the forecasted daily stock returns compared with the actual market returns. Both seem to have the ability to forecast direction, however they lack predicting the extent of daily return, with few unjustified peaks.

Even though the figures are doubtful, correlation coefficients suggest that Amazon is 6.5 times more predictable when compared with Apple. The 4^{th} hypothesis is accepted.

Lastly, this thesis could be a foundation for the study which would focus on stock returns forecasting of low capitalization such as non-dividend paying, young companies without financial history (which might be predicting the returns based on investor sentiments), investor interest and some other components. As both variables have improved its predicting significance with Amazon data. This research could guide investment banks or financial advisors investing in start-ups.

References

Baker, M., & Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *The journal of finance*, *61*(4), 1645-1680.

Baker, M., & Wurgler, J. (2007). Investor sentiment in the stock market. *Journal of Economic Perspectives*, *21*(2), 129-152.

Beardsworth , T. & Kumar, N. (2019). *Who to Sue When a Robot Loses Your Fortune*. Retrieved from https://www.bloomberg.com/news/articles/2019-05-06/who-to-sue-when-a-robot-loses-your-fortune?fbclid=IwAR38wpbdGVgFidnmR5YsvmaDkz_on9rzdcTtOv-YThQ6rURKO4e5CnFo558. [Accessed 14 May 2019]

Bloomberg. (2014). *Trending on Twitter: Social Sentiment Analytics*. Retrieved from https://www.bloomberg.com/company/announcements/trending-on-twitter-social-sentiment-analytics/. [Accessed 1 May 2019]

Bloomberg. (2016). *Trading the news: Use machine-readable data to find alpha*. Retrieved from https://www.bloomberg.com/professional/blog/trading-news-use-machine-readable-data-find-alpha/. [Accessed 1 May 2019]

Bloomberg. (2019). *A global icon*. Retrieved from https://www.bloomberg.com/professional/solution/bloomberg-terminal/?utm_source=bloomberg-facts. [Accessed 1 May 2019]

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of computational science*, *2*(1), 1-8.

Bordino, I., Battiston, S., Caldarelli, G., Cristelli, M., Ukkonen, A., & Weber, I. (2012). Web search queries can predict stock market volumes. *PloS one*, vol. 7, no. 7,

Browder, B. (2015). Red notice: a true story of high finance, murder, and one man's fight for justice. Simon and Schuster.

Brownlee, J. (2019). *A Gentle Introduction to the Rectified Linear Unit (ReLU) for Deep Learning Neural Networks*. Retrieved from https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks/. [Accessed 14 May 2019]

Campbell, J. Y., Grossman, S. J., & Wang, J. (1993). Trading volume and serial correlation in stock returns. *The Quarterly Journal of Economics*, *108*(4), 905-939.

Carrière-Swallow, Y., & Labbé, F. (2013). Nowcasting with Google Trends in an emerging market. *Journal of Forecasting*, *32*(4), 289-298.

Choi, H., & Varian, H. (2012). Predicting the present with Google Trends. *Economic Record*, *88*, 2-9.

Cont, R. (2001). Empirical properties of asset returns: stylized facts and statistical issues.

De Long, J. B., Shleifer, A., Summers, L. H., & Waldmann, R. J. (1990). Noise trader risk in financial markets. *Journal of political Economy*, *98*(4), 703-738.

Dertat, A. (2017). *Applied Deep Learning - Part 1: Artificial Neural Networks*. Retrieved from https://towardsdatascience.com/applied-deep-learning-part-1-artificial-neural-networks-d7834f67a4f6. [Accessed 14 May 2019]

Dimpfl, T., & Jank, S. (2016). Can internet search queries help to predict stock market volatility?. *European Financial Management*, *22*(2), 171-192.

Dugas, A. F., Jalalpour, M., Gel, Y., Levin, S., Torcaso, F., Igusa, T., & Rothman, R. E. (2013). *Influenza forecasting with Google flu trends*. PloS one, vol. 8, no. 2,

Dzielinski, M. (2012). Measuring economic uncertainty and its impact on the stock market. *Finance Research Letters*, *9*(3), 167-175.

Fama, E. F. (1960). *Efficient market hypothesis* (Doctoral dissertation, PhD Thesis, Ph. D. dissertation, University of Chicago Graduate School of Business).

Fleming, J., Kirby, C., & Ostdiek, B. (1998). Information and volatility linkages in the stock, bond, and money markets. *Journal of financial economics*, *49*(1), 111-137.

Griswold , A. & Karaian, J. (2018). *It took Amazon 14 years to make as much in net profit as it did last quarter*. Retrieved from https://qz.com/1196256/it-took-amazon-amzn-14-years-to-make-as-much-net-profit-as-it-did-in-the-fourth-quarter-of-2017/. [Accessed 15 May 2019]

Guresen, E., Kayakutlu, G., & Daim, T. U. (2011). Using artificial neural network models in stock market index prediction. *Expert Systems with Applications*, *38*(8), 10389-10397.

Hagstrom, R. G. (2013). *The Warren Buffett Way*, John Wiley & Sons.

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American economic review*, *93*(5), 1449-1475.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I* (pp. 99-127).

Lee, C. M., Shleifer, A., & Thaler, R. H. (1991). Investor sentiment and the closed-end fund puzzle. *The journal of finance*, *46*(1), 75-109.

Lemmon, M., & Portniaguina, E. (2006). Consumer confidence and asset prices: Some empirical evidence. *The Review of Financial Studies*, *19*(4), 1499-1529.

Li, X., Xie, H., Chen, L., Wang, J., & Deng, X. (2014). News impact on stock price return via sentiment analysis. *Knowledge-Based Systems*, *vol. 69*, pp. 14-23,

Liu, Y., Gopikrishnan, P., & Stanley, H. E. (1999). Statistical properties of the volatility of price fluctuations. *Physical review e*, vol. 60, no. 2,

Marr, B. (2018, A). *How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read*. Retrieved from https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/. [Accessed 7 May 2019]

Marr, B. (2018, B). *What Are Artificial Neural Networks - A Simple Explanation For Absolutely Anyone*. Retrieved from https://www.forbes.com/sites/bernardmarr/2018/09/24/what-are-artificial-neural-networks-a-simple-explanation-for-absolutely-anyone/#5cd7664b1245. [Accessed 14 May 2019]

Mi, L., & Hodgson, A. (2018). Real estate's information and volatility links with stock, bond and money markets. *Accounting & Finance*, *58*, 465. Retrieved from http://search.ebscohost.com.ludwig.lub.lu.se/login.aspx?direct=true&db=edb&AN=132515227&site=eds-live&scope=site [Accessed 12 May 2019]

Mian, G. M., & Sankaraguruswamy, S. (2012). Investor sentiment and stock market response to earnings news. *The Accounting Review*, *87*(4), 1357-1384.

Missing Link. (2018). *7 Types of Neural Network Activation Functions: How to Choose?*. Retrieved from https://missinglink.ai/guides/neural-network-concepts/7-types-neural-network-activation-functions-right/. [Accessed 14 May 2019]

Mittal, A., & Goel, A. (2012). Stock prediction using twitter sentiment analysis. Standford University, CS229 (2011 http://cs229. stanford. edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis. pdf), 15.

Neal, R., & Wheatley, S. M. (1998). Do measures of investor sentiment predict returns?. *Journal of Financial and Quantitative Analysis*, *33*(4), 523-547.

Oh, K. J., & Kim, K. J. (2002). Analyzing stock market tick data using piecewise nonlinear model. *Expert Systems with Applications*, *22*(3), 249-255.

Preis, T., Moat, H. S., & Stanley, H. E. (2013). Quantifying trading behavior in financial markets using Google Trends. *Scientific reports*, *3*, 1684.

Ramchand, L., & Susmel, R. (1998). Volatility and cross correlation across major stock markets. *Journal of Empirical Finance*, *5*(4), 397-416.

Rao, T., & Srivastava, S. (2012). Analyzing stock market movements using twitter sentiment analysis. In *Proceedings of the 2012 international conference on advances in social networks analysis and mining (ASONAM 2012)* (pp. 119-123). IEEE Computer Society.

Ripley, A. (2013). The smartest kids in the world: And how they got that way. Simon and Schuster.

Ryan, M. (2018). *A Simple Way to Know How Important Your Input is in Neural Network*. Retrieved from https://medium.com/datadriveninvestor/a-simple-way-to-know-how-important-your-input-is-in-neural-network-86cbae0d3689. [Accessed 15 May 2019]

Shiller, R. J. (1987). The volatility of stock market prices. *Science*, *235*(4784), 33-37.

Statcounter. (2019). *Search Engine Market Share United States Of America*. Retrieved from http://gs.statcounter.com/search-engine-market-share/all/united-states-of-america. [Accessed 10 May 2019]

Stephens-Davidowitz, S. & Pabon, A. (2017). Everybody Lies: Big data, new data, and what the internet can tell us about who we really are, New York, NY: HarperCollins

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of finance*, *62*(3), 1139-1168.

Wahome, R. (2018). *This Is How Twitter Sees The World : Sentiment Analysis Part One*. Retrieved from https://towardsdatascience.com/the-real-world-as-seen-on-twitter-sentiment-analysis-part-one-5ac2d06b63fb. [Accessed 10 May 2019]

Wanjawa, B. W., & Muchemi, L. (2014). ANN Model to Predict Stock Prices at Stock Exchange Markets. *arXiv preprint arXiv:1502.06434*.

Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., & Patwardhan, S. (2005). OpinionFinder: A system for subjectivity analysis. In *Proceedings of HLT/EMNLP 2005 Interactive Demonstrations*.

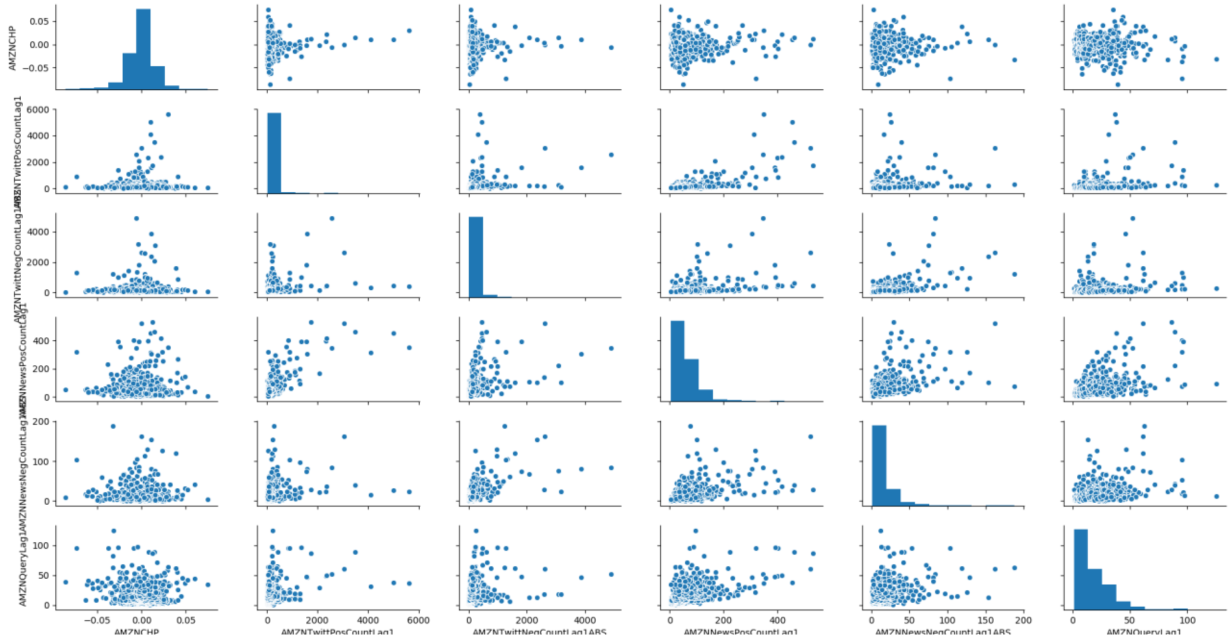
Wu, L., & Brynjolfsson, E. (2015). The future of prediction: How Google searches foreshadow housing prices and sales. In *Economic analysis of the digital economy* (pp. 89-118). University of Chicago Press.

Appendix A

Variables interactions: from the top (left) to bottom (right): Apple Daily Return, Apple Positive Twitter Sentiment, Apple Negative Twitter Sentiment, Apple Positive News Sentiment, Apple Negative News Sentiment and Investor Interest. Data Source: Google Trends and The Bloomberg Terminal, Note: Author's Calculation.

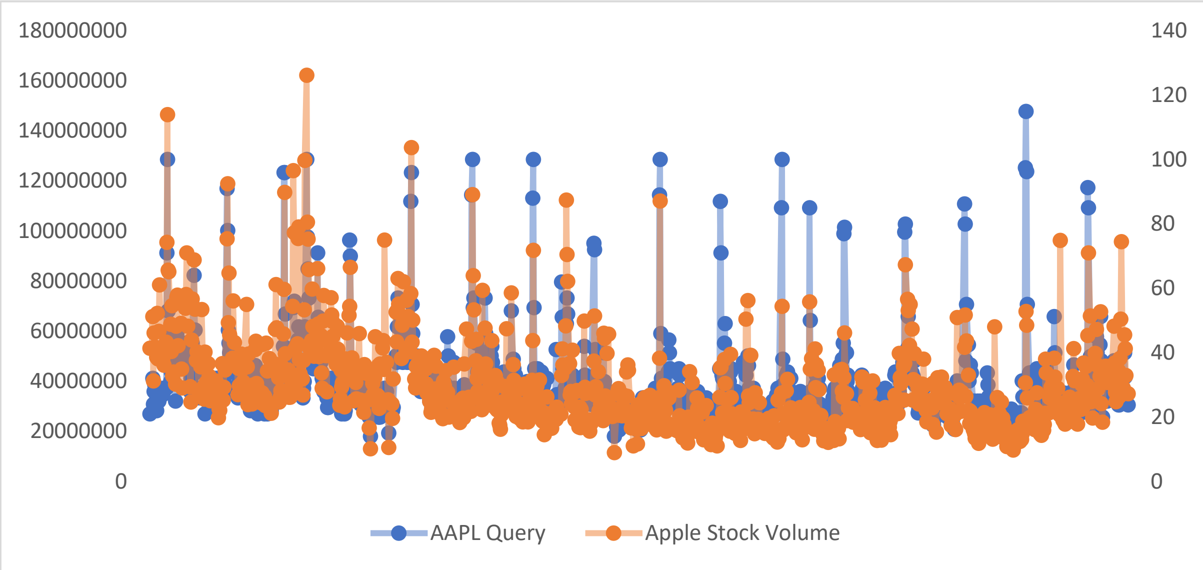
Appendix B

Variables interactions: from the top (left) to bottom (right): Amazon Daily Return, Amazon Positive Twitter Sentiment, Amazon Negative Twitter Sentiment, Amazon Positive News Sentiment, Amazon Negative News Sentiment and Investor Interest. Data Source: Google Trends and The Bloomberg Terminal, Note: Author’s Calculation.

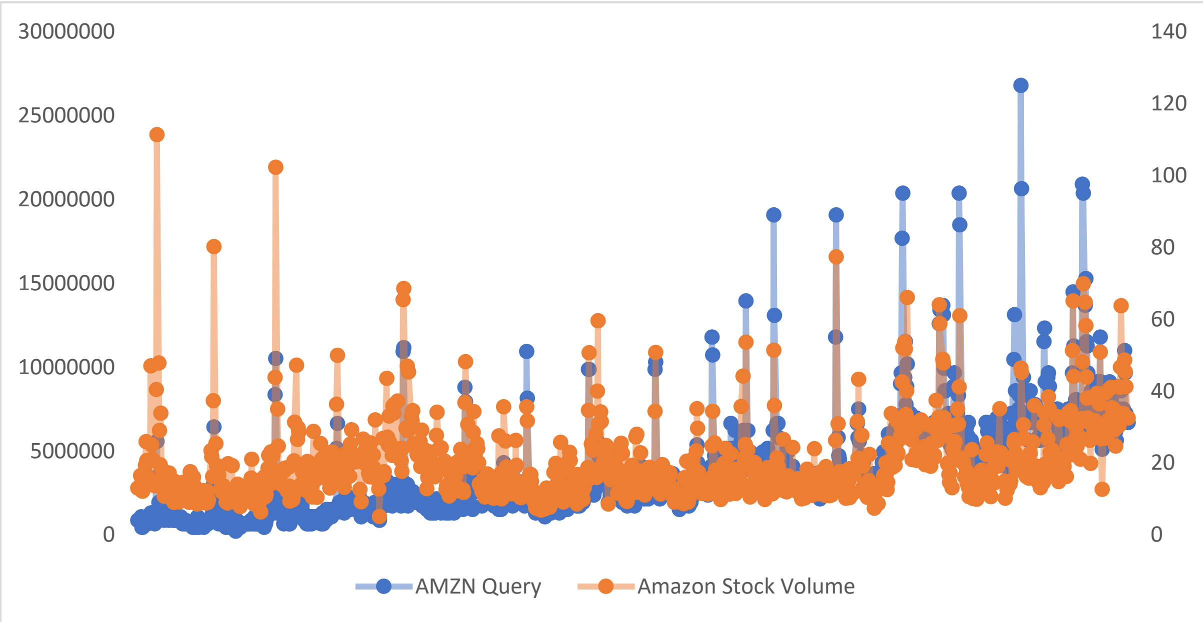


Appendix C

Correlation of Apple Stock Volume and AAPL Query (3 years), Data Source: Google Trends and The Bloomberg Terminal, Note: Author's Calculation.



Correlation of Amazon Stock Volume and AMZN Query (3 years), Data Source: Google Trends and The Bloomberg Terminal, Note: Author's Calculation.



Appendix D

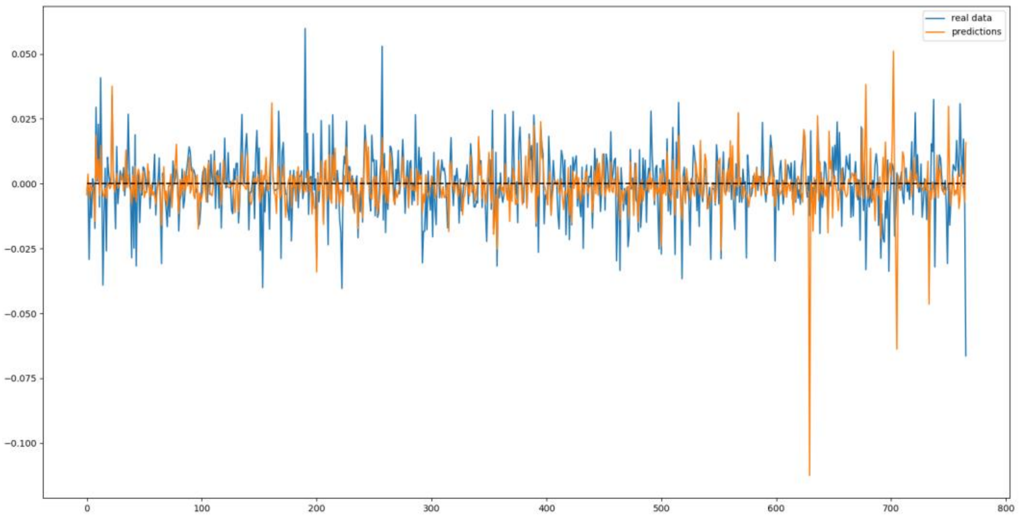
Source: Bordino, Battiston, Caldarelli, Cristelli, Ukkonen and Weber (2012)

Table 6. Values of cross-correlation functions for some selected stocks.

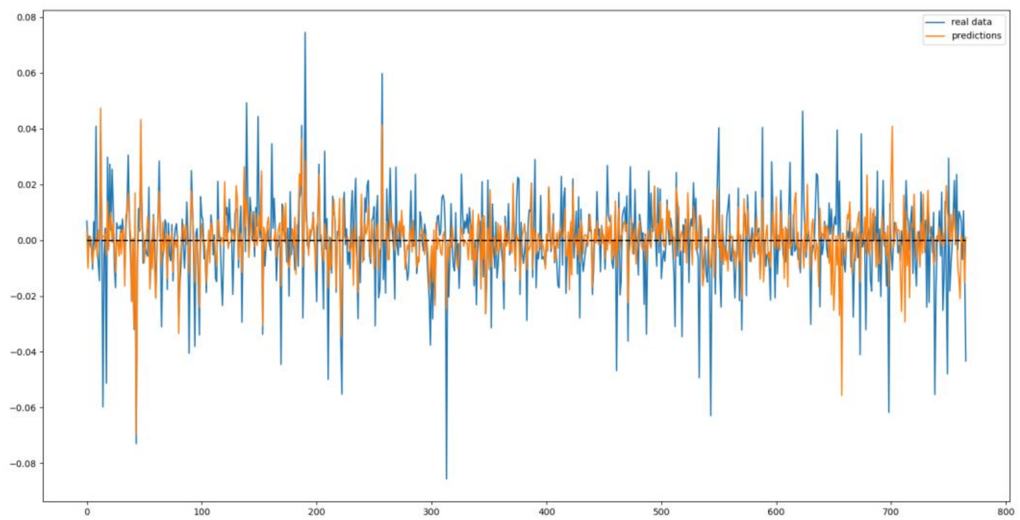
Ticker	$\delta=-5$	$\delta=-4$	$\delta=-3$	$\delta=-2$	$\delta=-1$	$\delta=0$	$\delta=1$	$\delta=2$	$\delta=3$	$\delta=4$	$\delta=5$
ADBE	0.08	0.12	0.14	0.19	0.47	0.83	0.51	0.19	0.09	0.10	0.11
CEPH	0.16	0.26	0.22	0.14	0.32	0.80	0.44	0.24	0.12	0.13	0.15
APOL	0.02	0.06	0.10	0.21	0.43	0.79	0.55	0.22	0.12	0.07	0.03
NVDA	0.23	0.36	0.38	0.46	0.56	0.79	0.68	0.47	0.42	0.38	0.29
CSCO	0.04	0.07	0.13	0.36	0.53	0.74	0.63	0.34	0.26	0.17	0.12
AKAM	−0.04	−0.06	0.03	0.07	0.22	0.72	0.49	0.20	0.11	0.02	-0.01
NFLX	0.10	0.16	0.16	0.24	0.47	0.68	0.54	0.25	0.19	0.16	0.13
ISRG	0.07	0.13	0.18	0.21	0.38	0.67	0.64	0.29	0.20	0.11	0.05
RIMM	0.03	0.12	0.11	0.14	0.31	0.66	0.58	0.24	0.20	0.11	0.05
FFIV	0.06	0.06	0.13	0.21	0.35	0.65	0.56	0.33	0.21	0.14	0.13

Appendix E

Apple training, Data Source: Google Trends and The Bloomberg Terminal, Note: Author's Calculation

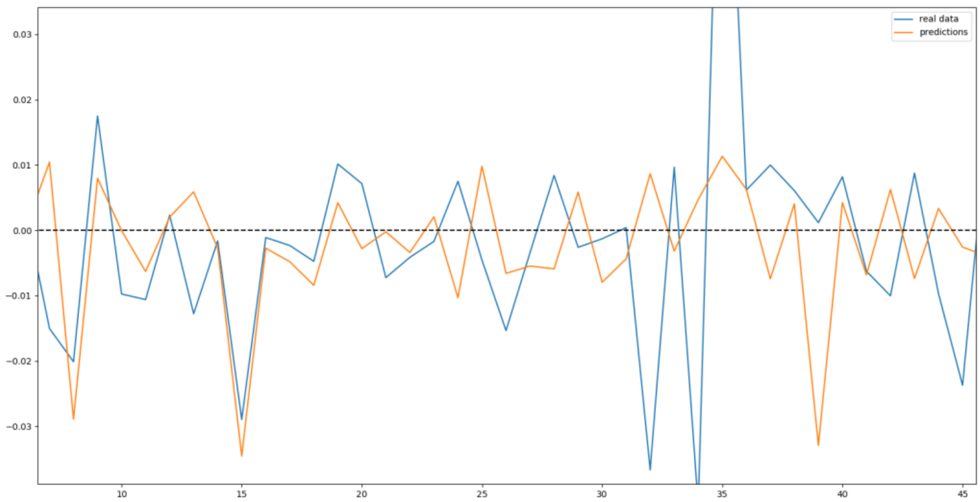


Amazon training, Data Source: Google Trends and The Bloomberg Terminal, Note: Author's Calculation



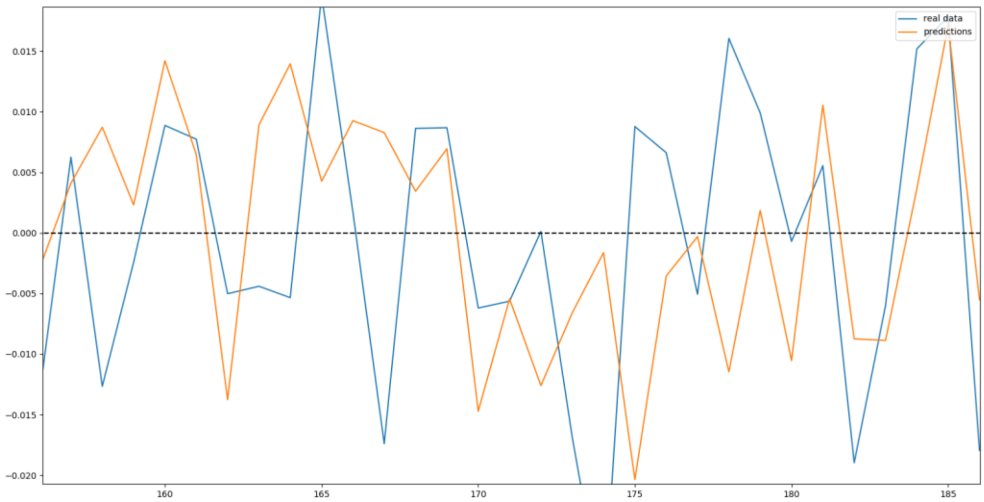
Appendix F

Apple Daily Return, ANN Model Predicting Values (Zoomed), Real Values in Blue and Predicted Values in Orange, Data Source: Google Trends and The Bloomberg Terminal, Note:



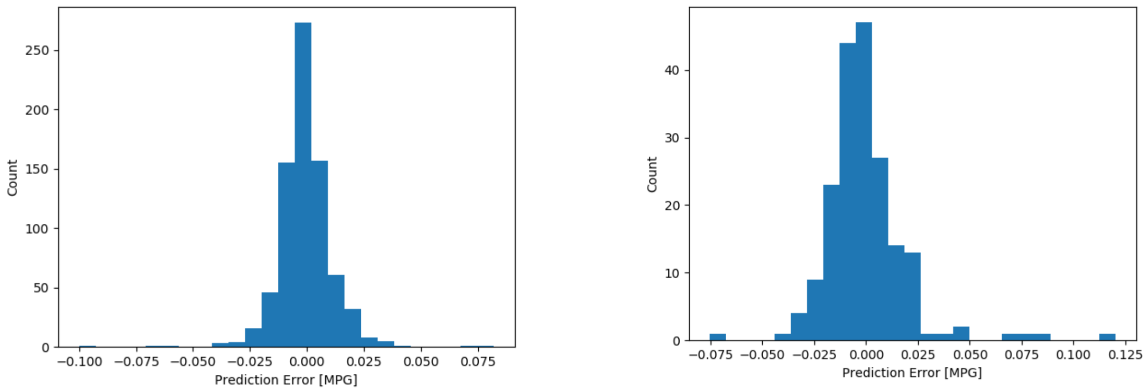
Author's Calculation.

Amazon Daily Return, ANN Model Predicting Values (Zoomed), Real Values in Blue and Predicted Values in Orange, Data Source: Google Trends and The Bloomberg Terminal, Note: Author's Calculation.



Appendix G

Frequency of prediction error of Apple, training (left) and testing (right), Data Source: Google Trends and The Bloomberg Terminal, Note: Author's Calculation



Frequency of prediction error of Amazon, training (left) and testing (right), Data Source: Google Trends and The Bloomberg Terminal, Note: Author's Calculation.