Twitter sentiment analysis using machine learning to predict volatility Index (VIX)

Zeddoug Youness, Sidibe Hamadoun, Mfumu-Kanda Dhi Kevin

May 2020

Abstract

This paper examines the relationship between Twitter sentiment related to daily financial events and volatility index over one week. The study contains twitter data from the big 5 tech company in the S&P 500 namely Amazon, Apple, Facebook, Google and Microsoft. Data from S&P 500 has been also included. A code in Python was written in order to collect data from Twitter and create a classification model using a Multilayer Perceptron. We find that our sentiment measure have a predictive power on the VIX index. Our trained model shows an accuracy of 66.66%.

1 Introduction

Trough various finance courses, we studied stock market's characteristics and more precisely, we explored the market efficiency theory (EMH) . According to this fundamental hypothesis, we know that stock prices are driven by new information. Hence, we asked ourselves : which information precisely ? What type of information exactly ? through press release or informal channel (social media) ? We investigated several sources of information such as company website and finally we focused on Twitter. Indeed, this social media allows users to express ideas openly with limited words. Due to this limitation, individuals condense their ideas and provide only relevant information. Moreover, users feel at ease to express their emotions and reactions about any events. Our choice to select Twitter is supported by Nofsinger's paper (2005). With significant and consistent results, this study highlights that decisions and more precisely financial decisions are driven by emotions and mood. For many years, we are familiar with the massive growth of social media around the world. According to Statistica and TNW, (2019), it is over than 2.2 billions of users which are browsing on Facebook or even 329.50 millions users on Twitter. We observed the consistent and efficient influence of social media within modern society particularly in Politics, Science and obviously in Finance. In 2016, Trump became the 45th president of the USA. His excessive use of Twitter confirms that we entered in a new era of persuasion and communication. Indeed, it is very novel

and informal for a president to express emotions and ideas in a spontaneous way through social media. However, Twitter does not only serve the president of the USA. Indeed, Twitter is a large and developed platform that allows users to interact by expressing personal or professional opinions for instance press release. Moreover, as Pikulina, Renneboog and Tobler (2017) argue, social media became a gold mine for financial data and data analysis purpose. It allows us to explore the corresponding literature about the impact of individual's sentiment on financial dynamics such as volatility.

2 Description of the research question and the relevant literature

In this study, we investigate the potentially effects of individuals' sentiment concerning financial events on the volatility index (VIX). In this part : first, we provide succinct definitions of sentiment analysis and microblogging. Second, we explore the relevant literature trough these two concepts and obviously their implication on stock dynamics.

2.1 Definitions

According to the Oxford dictionary, sentiment analysis is defined as the "process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, event is positive, negative or neutral".
By carefully exploring scientist's work in Finance or Computer Science, we discovered a recent subject : microblogging. Java et al. (2007) provide a comprehensive definition of this "trendy" concept : "microblogging is a new form of communication that allows you write concise text updates in which users can express their current status in short posts distributed by instant messages, mobile phone, email or the web". Further, we will notice that these two fundamental subjects are absolutely connected.

2.2 Literature review

With the sustainable growth of medias and social networks, researchers extend their research domain and tend to explain the implication of these platforms on finance and data mining. By exploring the existing literature, we noticed several techniques that practitioners and scientists employ for text's analyzes. As first sight, we identified a certain chronology in the individual's attitude analysis topic. One of the first paper which propose to recognize individual's intentions trough text structure was written by Grosz in 1979. This study highlights the fact that somebody extract from a particular set of words an interpretation of what he wants to describe. It is a major debut for the computational aspects of linguistic structures literature. Indeed, it is only an introduction of a deep pathway concerning the sentiment analysis theory. Then, professor Antonakis from

the University of Lausanne published several papers about evaluating charisma and sentiment to some extent trough a written or oral speech, (Antonakis, Fenley and Liechti, 2011), (Antonakis, 2012). It is curious to visualize the continuous evolution of capturing clue, information and individual's attention through a set of words. Progressively, with the emergence of advanced algorithms and elaborated machine learning theory, researchers initiated diverse sophisticated techniques. Indeed, Pak and Paroubek, (2010) provide an investigative approach for sentiment analysis. In fact, they decided to inspect emoticons included in a text. In this respect, Pak and Paroubek, (2010) divided into three classes emoticons' impression : positive, negative, neutral. For instance, a symbol corresponding to ":)" is classify as a positive emotion. In order to proceed and provide results, Pak and Paroubek, (2010) implemented the Naïve Bayes classifier. As a result, they could conclude about the communicative intention of an individual. Also, we examined the Stocktwits application (a social service specifically targeted for investors and traders) and we also remarked that users' tweets were classified into three units : bearish, bullish and neutral. These classifiers propose an overview of a user's sentiment about a specific event, a company or a decision. Besides, according to Oliveira, Cortez and Areal, (2013), institutions as the American Association of Individuals Investors propose actually several surveys containing sentiment classifiers.

Actually, we noticed that it is very intuitive to analyze individuals' attitude and opinions towards a subject through microblogging platforms such as Twitter.Indeed, individuals use these platforms because it is more convenient to express ideas in substance and form. In the sense that, the written form and style is unrestrained. Moreover, in Twitter for instance, we can easily evaluate the influence and the impact of a tweet in the community through number of retweets and likes.In other words, we can quantify the influence of individuals' statements. According to Java et al. (2007), we observe that these microblogs allow an instantaneous and prompt communication. Indeed, extracting data from microblogging platform simplify financial dynamics' prediction. For instance, an index is updated instantaneously; considering new information every second from microblog constitute a relevant asset in order to predict volatility. Considering this fundamental setting, we expect some robust evidence in the volatility index prediction.

Hence, According to Souza et.al (2015), twitter as other microblogging platforms have become relevant sources to describe financial dynamics such as stock prices and volatility. Moreover, everyone has the opportunity to express his opinion. Indeed, from profane to experts in Finance, each user can share, influence the general tendency with his own words. According to U.S Securities and Exchange Commission report (2013), companies were allowed to publish relevant information concerning them on Twitter. It clearly shows the primordial benefit of microblog in firms' announcement nowadays[retravailler la phrase]. As an illustration of recent impact of microblogging on financial dynamics in stock market, we want to mention the story about the hacked Twitter account of the American news agency Associated Press. In 2013, someone used this fake

account in order to announce a possible attack on the White House. According to the Wall Street Journal (2013), this false disclosure caused a drop in the Dow Jones Industrial Average of 145 points in minutes.

It allows us to present several empirical evidence of the human's sentiment impact on stock dynamic. Siegel, (1992) already states with interesting results that a change in individuals' sentiment towards an event generate a volatile environment for stock prices. To some extent, Lee, Jiang and Indro, (2002) argue that a transformation in bullish (bearish) attitude conduct to a downward revision (upward) of volatility in stock market.

3 Methodology

3.1 Proposed method

In order to provide an explanation for this research question, we propose a method consisting of 3 phases:

The first phase relates to the initial extraction and treatment of both financial and twitter data.

For the second phase, we work with NLTK's tool VADER. This phase concerns both the pre-processing and the sentiment attribution of the dataset. These internal intermediate steps yield respectively to so-called tokens and to a sentiment classification on a continuous scale from 1 to -1, depending on the sentiment.

Finally, in the last part, we opted for a multi-layer perceptron method in order to extract financial meaning from the obtained sentiment. This allowed us to determine whether a change in sentiment can trigger a change in price movements of the VIX volatility index.

3.2 Data Collection

3.2.1 Twitter data collection

Collecting twitter data is a crucial step to be able to empirically detect a pattern between financial markets movements and the behaviours of its participants. Generally, Twitter data collection is made possible by the platform's API, which has evolved over time to also offer sentiment analysis features ("Analyze – Twitter Developers", 2020). Other commercial platforms like StockTwits, also offers access to specific Twitter data.

Nevertheless, their access has gradually become limited to paid members making it impossible for us to fully access them. We therefore used the standard version of Twitter API, freely accessible to Twitter developers, in order to retrieve the maximum authorized number of relevant data, a sampling of the recent Tweets published in the past 7 days.

3.2.2 Financial data collection

The original idea was to consider 30 blue chips companies in the US stock index, the Standard Poor 500 (S&P 500), since other securities might not be covered as well in terms of tweets. But for reasons explained in the previous section, we had to reduce this number. Therefore, this paper focuses on the five biggest constituents of the S&P 500 in terms of market capitalization, namely Apple (ticker: AAPL), Amazon (ticker: AMZN), Facebook (ticker: FB), Microsoft (ticker: MSFT), Google (ticker: GOOG) and on our financial indicator of interest the volatility index (ticker: VIX). We used the Chicagor Board Options Exchange (CBOE) platform to retrieve the data for the stocks and the indices prices.

3.2.3 Choice of time period

Concerning the time period, the dataset goes from of the 15 of May 2020 to the 22 of May 2020.

3.2.4 Choice of financial platforms

The VIX indicator prices were obtained from Yahoo! Finance. We considered daily closing prices for our analysis, as well as the corresponding volumes.

3.3 Pre-processing

Micro-blogging data like the ones from Twitter are inherently hard to draw features and patterns from because of their imprecision. Another layer of complexity is added when these data are industry-specific as their meaning might vary greatly from the general language one. In light of these elements, textual data have to be transformed into a format that captures their meaning and their information in an unambiguous manner. That is the role of pre-processing. In this section, we introduce some of the main pre-processing methods applied to textual data, before discussing the one applied in our analysis.

3.3.1 Tokenization

This is usually the first step in NLP applications when working with textual data. Tokenizing the data means breaking them up into units called tokens, which are generally the smallest useful unit for semantic processing. However, tokens might also be punctuation marks or even numbers. Additionally to the splitting of sentences into words, a number of tokens are also removed as part of this tokenization process. These are punctuation marks and words that appear extremely frequently like “a”, “and” or “but”. These latter are called stopwords.

Another typical phase of the tokenization process is called Stemming and Lemmatisation, which are actually two processes that are often used in conjunction. According to Guida, (2019), their objective is to reduce words from their

derived grammatical forms to a reduced form, the so-called base form or root form.

The objective behind stemming is to group related words into the same stem. It does work even if the system generated stem is not a valid root. Some of the most famous stemming algorithms have been developed decades ago by Guida, (2019) and Lovins, (1971). It is a relatively simple tool to implement, a simplicity that is also a disadvantage as many words are transformed into erroneous roots (Lovins, 1968). Lemmatisation consists of finding the dictionary form of a given word. For example, cars become car and replay stays replay. In English, both methods yield the same root but they actually are different.

In theory, these methods seem simple, and algorithmically some of them are but their actual implementation is particularly difficult as word inflections have a lot of irregularities that might require the development of separate rules for each of them, which becomes very tedious (Heidenreich, 2018).

Other methods exist to treat textual data like vocabulary-based ones, where the approach is to target a few specific words or phrases among others in a given text, or Part-of-speech tagging, which is the process of assigning a token according to the grammatical category (Ramachandran, 2018).

3.4 Token into features

After having retrieved the smallest useful unit as tokens, the next phase consists of making sense of them, of actually conducting a sentiment analysis. It is particularly important as most of the news or tweets are produced for human consumption and are not fit for computer analysis.

3.4.1 Bag of words

The most commonly used method is called the bag of words. It consists of encoding a document as an unordered list of the words it contains, generally without paying too much attention to context and grammar, only multiplicity is retained (Guida, 2019). Here the objective is to measure the frequency of each word and record it in a vector/ matrix. In this case again, there exists several ways to represent the bag of words. One of the most common one consists of rows that represent individual words and columns that provide the word counts per document (Guida, 2019). This method has obviously a lot of disadvantages. It does not consider the order meaning. Considering this order meaning is one of the extensions of the bag of words method, it is called the N-gram method, which basically parses the words into a sequence of 2 or even sometimes 3 words. Theoretically, the larger the n the more meaning the model can store. However, this method does come as well with a few flaws like the possible infinite number of pairs that are possibly generated. More filtering still needs to be done. At this point, it is also possible to understand more complex data by using neural networks.

3.5 Sentiment analysis

For this analysis we opted for Python's Natural Language toolkit (NLTK) that makes the pre-processing of textual data relatively easy. We converted the tweets to lowercase, stemmed them by using an NLTK sub-package called "Porter stemmer" and tokenize them into individual words. To capture the sentiment from these tokens, we used another NLTK sub-package called The Valence Aware Dictionary and sentiment Reasoner or VADER. It basically is a pre-trained algorithm that uses a simple rule-based model to generate sentiment for analysis. Its algorithm is optimized for social media like Twitter (Genç, 2019). Practically, Vader is a lexicon or dictionary of sentiment containing current words and expressions used in the popular culture. We used this method to obtain a sentiment score that could be standardized in a range of -1 to 1.

3.6 Classification

The last phase of our analysis is the treatment of the sentiment obtained in the previous section. In order to do so, we decided to use a simple multi-layer perceptron or neural network.

The perceptron is an algorithm used in machine-learning, more particularly for supervised learning of binary classifiers. These latter decide whether a given input belongs to a specific class or not. There exist two types of perceptron, the single-layer perceptron, which can be considered the basic unit of a neural network and the multi-layer perceptron (MLP), which is basically several single-layer perceptron stacked together. The Perceptron algorithm takes weighted inputs, processes them and is capable of performing binary classifications. The MLP primary purpose is to create a model that can solve complex computational problems from large sets of data. It is composed of the four following parts:

3.6.1 The input values

The perceptron takes real values as inputs. In our case, we fed the model with the sentiment obtained in the previous phase.

3.6.2 The weights and their sum

The weights allow the model to determine the relative importance of each of the inputs in the output obtained. Assigned randomly at first, they are fine tuned in an iterative process called back-propagation [30]. Each time the weighted sum of these inputs is computed.

3.6.3 The activation function

The activation function takes the weighted sum and apply a step rule function to it. This function has several characteristics, it converts the numerical output into a binary output of generally -1 and 1 or 0 and 1, helps the perceptron learn

when part of a MLP and makes it possible to train the model in a non-linear way ("What is Perceptron — Simplilearn", 2020).

3.6.4 The perceptron output

The output is a classification decision of the given inputs. In the case of a multi-layer perceptron, the output of one layer is the input of the next one but the final classification decision happens at the final layer ("Perceptrons Multi-Layer Perceptrons: the Artificial Neuron - MissingLink", n.d.).

4 Results

This section discusses the two main results found during the implementation phase. We will first look at the sentiment score obtained. Then we will analyse and comment the performances of the Multilayer Percetron.

4.1 Sentiments results

Sentiment score are a key ingredient in this paper as the objective is to investigate if the return of the implied volatility can be predicted by using tweets sentiments. In order to get these scores we used the VADER sentiment analyzer which has already been explained in section 3.5. Figure 1 shows six business days sentiment level for each hashtag. During this period, we can notice a general positive sentiment. Only google shows high volatility. The company register the lowest sentiment rate on may 21, the day after they announced the release of their coronavirus-tracking app in collaboration with Apple. At the same date we also observe the highest score of the sample made by Apple. These results suggest that the users of twitter judged that this new collaboration will have different impact on the two concerned companies.

In addition, the sentiment level is higher for the individual stocks than for the SP500. A potential explanation for this observation could be the fact that SP500 is an index and therefore represents an aggregate of sentiment of all its components.

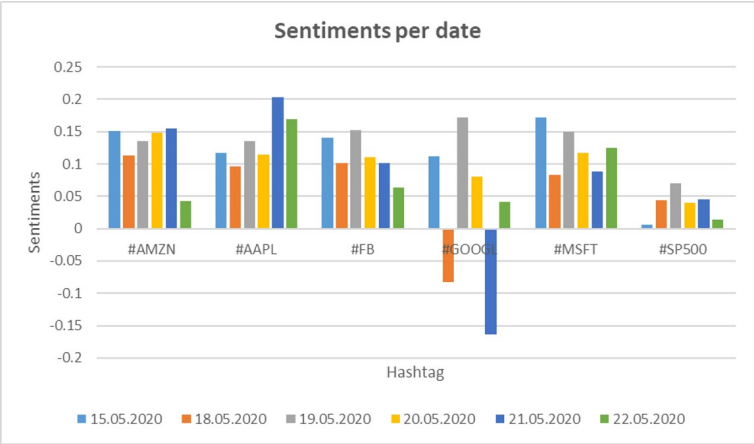

Figure 1: Daily sentiment score per Hashtag

4.2 MLP results

This part of the section presents the results drawn from the Multilayer Perceptron used as model to classify the return of the VIX index. The figure 2 shows the architecture used for this purpose. All the implementation has been made on python with the Keras library. To avoid over-fitting issue due to small data set we decided to only use one hidden layer. The initial layer composed of 7 nodes. A vector of six elements plus the bias.

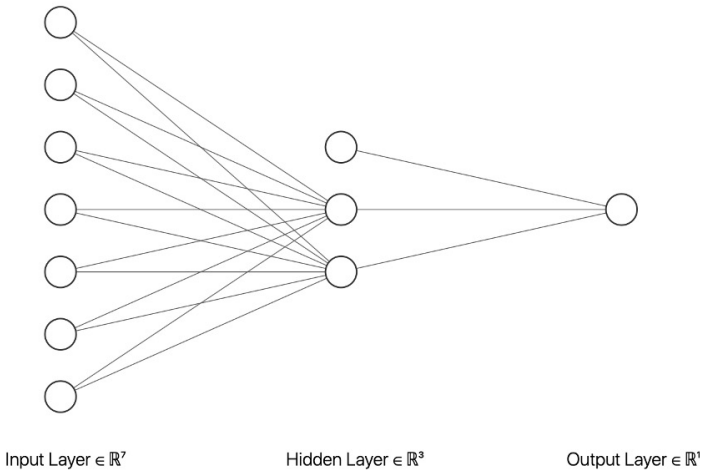

Figure 2: Neural network architecture

At a first glance the results found seem quite interesting even with a very small set of data. On figure 3 we can see a summary table of the performances.

On one hand we have the value of the loss function and on the other we have the accuracy metric scores. Notice that here the accuracy score is binary. Either 0 or 1. The reason coming from the fact that we are using 1/6 of data as test set. As the total data set has a length of 6 this means that for each cross-validation we only predict one output. Therefore, if this prediction is correctly classify, we get a 100% accuracy otherwise 0%.

Out of the 6 predictions, the model successfully classified 4 correctly. This give us an average accuracy of 66.667%. If the analysis was only based on these accuracy scores, we could conclude that we have a good model because we can predict more than 50% of the time the direction on which the VIX future will go. Unfortunately, this is not entirely the case. Indeed, by taking into account the loss which has a mean of almost 1, we can see that there is still room for improvement.

A reflection that one can have is on the significance of the results. The neural network presented in this paper has been trained and tested with only a every small set of data. To obtain a robust model it is crucial to use a larger amount of data. Using large date set may helped building a consistent model on which one would probably rely on to take strategic or tactical decisions.

Moreover, even if the results suggest that volatility can be predicted with tweets sentiments, that does not mean that a simple trading strategy such as buying the VIX future when you have a bullish signal and the selling it when it's a bearish signal will be sufficient in order to take advantage of this predictability. One should also consider the empirical performance of this strategy. It could be the case that the magnitude of your losses when the model predicts the wrong direction exceed by far more the gains earned by a correct prediction.

	Loss	**Accuracy**
Mean	**0,9941**	**0,6667**
	0,3836	1
	0,1813	1
	2,8287	0
	0,4988	1
	1,8645	0
	0,2077	1

Figure 3: Cross validation loss and accuracy score table

5 Discussion and limitations

In this section, we essentially discuss several limitations concerning our results and how we proceeded during this research. First of all, we highlighted in this project the key role of data collection and how it is challenging to obtain a sufficient number of relevant observations. Nevertheless, for access rights' purpose we did not have the opportunity to collect enough data in order to say that our results are robust. Indeed, we loose in terms of robustness and obviously, we should interpret thoroughly our conclusion for this project. In other terms, we can say that results can suffer from bias selection due to the insufficient numbers of observations.

Moreover, we noticed the fact that a specific setting in this paper could be improved. Indeed, we refer to the sentiment analyzer's tool (i.e : VADER) which is a dictionary of sentiment containing words and expressions largely used. However, the limitation is the following : in this project, we consider financial words and technical expressions too. Perhaps, this tool does not fully capture the hidden meaning of some specific financial industry expressions economic expressions and obviously reduce the implication of such words in the final result. For instance, an expression like "bull market" will be classify as a positive signal in the financial industry while this same expression is only consider as a combination of two standards words in a general speaking. In order to rectify this limitation, we can advice to implement a neural network and train it on financial tweets data. The later will help us a more precise sentiment analyzer.

After these technical considerations, we also observed two possible issues concerning the temporal aspects between financial markets and tweets. The first one being that tweets could be considered as an impulsive behaviour in a financial theory sense, and as such the sentiments they reflect may greatly vary and at a high frequency than the daily time period considered for our financial data. It becomes problematic once we assume that these tweets might indeed trigger market movements or volatility as considered in this paper. Simply put, the daily volatility (aggregate of the volatility in a sense) we considered might be very different from the intra-day volatility triggered by the continuous flow of tweets. The second issue we observed is that the tweets are continuous, whereas the stock markets are open only on week days. This discrepancy should also be considered for in some way in a future analysis.

Finally, we can discuss the classification method used to tackle this prediction question. Indeed, we decided to implement a Multilayer Perceptron which provides an accuracy score of 66.667%. Due to the weak number of observations, we think that a neural network may not an appropriate model to use. As a result, we decided to explore an other interesting method : support vector machine (i.e SVM). With the latter classifier we obtained the same accuracy score : 66.667% as the MLP. The SVM is less costly in term of implementation,

hence we think it can be probably more convenient to use this method in this case.

6 Conclusion

blabla

References

[1] *Analyze – Twitter Developers. Developer.twitter.com. (2020). Retrieved 20 May 2020, from https://developer.twitter.com/en/use-cases/analyze.*

[2] Guida, T. (2019) *Big data and machine learning in quantitative investment (1st ed., pp. 567-600). Wiley.*

[3] Lovins, J. (1971) *Error Evaluation for Stemming Algorithms as Clustering Algorithms [archive], JASIS, 22: 28–40*

[4] Lovins, J. B. *"Development of a Stemming Algorithm." Mechanical Translation and Computational Linguistics 11, 1968, 22—31*

[5] Heidenreich, H. (2018) *S*temming? Lemmatization? What?. Medium. Retrieved 22 May 2020, from https://towardsdatascience.com/stemming-lemmatization-what-ba782b7c0bd8.

[6] *Ramachandran, A. (2018). NLP Guide: Identifying Part of Speech Tags using Conditional Random Fields. Medium. Retrieved 22 May 2020, from https://medium.com/analytics-vidhya/pos-tagging-using-conditional-random-fields-92077e5eaa31.*

[7] *Big data and machine learning in quantitative investment (1st ed., pp. 567-600). Wiley.*

[8] Genç, Ö. (2019) *T*he basics of NLP and real time sentiment analysis with open source tools. Medium. Retrieved 22 May 2020, from https://towardsdatascience.com/real-time-sentiment-analysis-on-social-media-with-open-source-tools-f864ca239afe.

[9] Nofsinger, J. R. (2005). *Social mood and financial economics. The Journal of Behavioral Finance, 6(3), 144-160.*

[10] https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

[11] Pikulina, E., Renneboog, L., & Tobler, P. N. (2017). *Overconfidence and investment: An experimental approach. Journal of Corporate Finance, 43, 175-192.*

[12] Java, A., Song, X., Finin, T., & Tseng, B. (2007, August). *Why we twitter: understanding microblogging usage and communities. In Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis (pp. 56-65).*

[13] Grosz, B. J. (1979). *Focusing and description in natural language dialogues (No. SRI-TN-185). SRI INTERNATIONAL MENLO PARK CA.*

[14] Antonakis, J., Fenley, M., & Liechti, S. (2011). *Can charisma be taught? Tests of two interventions. Academy of Management Learning & Education, 10(3), 374-396.*

[15] Antonakis, J. (2012). *Transformational and charismatic leadership. The nature of leadership, 256-288.*

[16] Pak, A., & Paroubek, P. (2010, May). *Twitter as a corpus for sentiment analysis and opinion mining. In LREc (Vol. 10, No. 2010, pp. 1320-1326).*

[17] Oliveira, N., Cortez, P., & Areal, N. (2013, September). *On the predictability of stock market behavior using stocktwits sentiment and posting volume. In Portuguese conference on artificial intelligence (pp. 355-365). Springer, Berlin, Heidelberg.*

[18] Souza, T. T. P., Kolchyna, O., Treleaven, P. C., & Aste, T. (2015). *Twitter sentiment analysis applied to finance: A case study in the retail industry. arXiv preprint arXiv:1507.00784.*

[19] Lauricella, T., Stewart, C., Ovide, S. (2013). *Twitter Hoax Sparks Swift Stock Swoon. WSJ. Retrieved 18 May 2020, from https://www.wsj.com/articles/SB10001424127887323735604578441201605193488*

[20] Siegel, J. J. (1992). *Equity risk premia, corporate profit forecasts, and investor sentiment around the stock crash of October 1987. Journal of Business, 557-570.*

[21] Lee, W. Y., Jiang, C. X., & Indro, D. C. (2002). *Stock market volatility, excess returns, and the role of investor sentiment. Journal of banking & Finance, 26(12), 2277-2299.*

[22] *Backpropagation. DeepAI. (2020). Retrieved 25 May 2020, from https://deepai.org/machine-learning-glossary-and-terms/backpropagation.*

[23] *What is Perceptron — Simplilearn. Simplilearn.com. (2020). Retrieved 25 May 2020, from https://www.simplilearn.com/what-is-perceptron-tutorial.*

[24] *Perceptrons Multi-Layer Perceptrons: the Artificial Neuron - MissingLink. MissingLink.ai. Retrieved 25 May 2020, from https://missinglink.ai/guides/neural-network-concepts/perceptrons-and-multi-layer-perceptrons-the-artificial-neuron-at-the-core-of-deep-learning/.*

[25] Press, A. (2020). *Apple, Google release technology for coronavirus-tracking apps. MarketWatch. Retrieved 24 May 2020, from https://www.marketwatch.com/story/apple-google-release-technology-for-coronavirus-tracking-apps-2020-05-20?mod=mw_quote_news*