Analysing Tweets Sentiments for Investment Decisions in the Stock Market

Citation for published version:
Hao, Z & Chen-Burger, Y-HJ 2021, Analysing Tweets Sentiments for Investment Decisions in the Stock Market. in *Agents and Multi-Agent Systems: Technologies and Applications 2021.* Smart Innovation, Systems and Technologies, vol. 241, Springer, pp. 129-141, 15th International KES Conference on Agent and Multi-Agent Systems-Technologies and Applications 2021, Virtual, Online, 14/06/21. https://doi.org/10.1007/978-981-16-2994-5_11

Digital Object Identifier (DOI):
10.1007/978-981-16-2994-5_11

Link:
Link to publication record in Heriot-Watt Research Portal

Document Version:
Peer reviewed version

Published In:
Agents and Multi-Agent Systems

Sentiment Analysing Tweets for Stock Market Movements

Zhicheng Hao and Yun-Heh Jessica Chen-Burger

Heriot-Watt University, School of Mathematical & Computer Sciences
Edinburgh, UK, zh3@hw.ac.uk, y.j.chenburger@hw.ac.uk

Abstract. The common practice of the public for using social media as information sources and basis for decision making has made the social media an important alternative media. This is in particular true for investors in the stock market due to their needs to gain real-time information and strategic persons' opinions. It is therefore very interesting to investigate the relationships between movements and trends of stock market prices and text as published in the social media platforms. In this paper, we analyse text published in Twitter to identify their sentiments and use this information to determine whether there are relationships to the stock market price movements or not. Sentiment Analysis derives opinions from people who created content on social media platforms. This study analyses Tweets from Twitter users who are influential in the financial sector (inc. Bloomberg, Reuters, Donald Trump, Forbes, Wall Street Journal) and compare the sentiments of their Tweets. In this paper, the Python tool Twint is used to scrape Tweets. SentiStrength, a lexicon-based sentiment analysis tool, is used to indicate the polarity and sentiments of the text. The collected Tweets are labelled with positive, neutral and negative scores. These scores are then combined to generate composite scores. These are then compared with the manually labelled sentiment scores (ground truth) to analyse the performance of the tool. Multiple evaluation metrics are used, inc. accuracy, precision, recall, and F1 to provide additional performance indications.

Keywords: Sentiment Analysis, Twitter, Stock Market Price, SentiStrength.

1 Introduction

The financial-market-based information is increasingly attracting attention in recent years. Traditionally, investors heavily depend on information reported in financial news to decide whether to buy, sell or hold stocks in the stock market. These days, social media provides much speedier, near real-time market information and from rich sources of information, inc. independent investment advisers, personality's self-publishing, or government's financial announcements. One such example is Tweets as published in Twitter. Twitter users produce millions of Tweets simultaneously which can be used to gauge stock market sentiments and thus to analyze and predict stock market movements. However, it is difficult to predict the financial market accurately using Twitter information. Because such Tweets may be complex that they include additional embedded information; they may also include unrelated or ambiguous information, such as sarcasm. Carefully designed Sentiment Analysis tools can be designed to remove noises and extract relevant features from them. There are several useful tools that can be used to carry out sentiment analysis, such as SentiStrength, Weka, and NLTK.

Over the years, microblogging platforms have become one of the major sources of information, inc. Twitter, Facebook, and Instagram. Companies and organizations are increasingly seeking methods to explore social media for information about what people think about their services and products. Twitter is such a micro-blogging social media platform that allows users to send and read a message of up to 140 characters. Twitter has a large user base globally and is a great narrative of public mood. It therefore also has great potential for exploration in the stock market. Twitter provides a huge amount of information that can be extracted to provide features to analyze and mine more hidden information. To gauge public mood, the traditional way of survey by questionnaires can be very expensive and time-consuming to investigate sufficiently large samples [1]. However, Twitter includes a lot of topics which may include unrelated, un-important and/or ambiguous information. It can be challenging to identify relevant, accurate, unambiguous and key information from Twitter.

Text mining and sentiment analysis techniques have been used to analyze mood in blogs [2]. However, performing sentiment analysis on Tweet messages is difficult due to the widely use of informal languages and expressions, inc. slang, abbreviations, icons, misspells, ambiguities and sarcasms. The sentiment analysis techniques can be divided into two groups: the first group includes the use of lexicon of positive and negative words [3]. The second group uses machine learning classifier such as Support Vector Machine and Liner regression [4]. There are several useful sentiment analysis tools such as Weka, SentiStrength, and Mozdeh. Investors and researchers use these tools for analysis to the price discovery process and to make smarted invest decisions. Sentiment analysis can judge opinions from Twitter as positive, negative, and neutral [5]. In addition, sentiment analysis tools can be used to find polarity of the text. For example, SentiStrength reports two sentiments which are -1 (not negative) to -5 (extremely negative) and 1 (not positive) to 5 (extremely positive).

True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN) are used to evaluate the accuracy between predictive sentiment and actual sentiment [6]. True Positive (TP) is text which both predictive and actual sentiment are positive. False Positive (FP) is text that is wrongly predicted as positive when the actual sentiment is negative. True Negative (TN) is text which both predictive and actual sentiment are Negative. False Negative (FN) is text that is wrongly predicted as negative when the actual sentiment is positive. Table 1 shows relationships between predictive and actual results.

Table 1. Relationship between predicted results and actual results

	Predicted Results	
	Positive	Negative

Actual Results	Positive	True Positive (TP)	False Negative (FP)
	Negative	False Positive (FP)	Ture Negative (TN)

Additional evaluation measures can be calculated based on these four variables such as accuracy, precision, recall, and F1 Score. [6] Accuracy is an important measurement of performance. Accuracy is the number of predicted results correctly divided on the overall results. The accuracy can be defined as follows:

$$Accuracy = \frac{Numbers\ of\ results\ predicted\ correctly}{Counts\ of\ all\ possible\ results} \quad (1)$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (2)$$

There are other widely used metrics for evaluating performance such as precision, recall and F1 score. Precision is used to measure exactness and recall measures completeness. Precision is the number of results correctly predicted divided by the overall number of predicted positive results. Recall is the number of results correctly divide by the overall number of results that are truly are positive.

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

In addition, F1 score is a balance between precision and recall. F1 score is the harmonic mean of precision. When the result is difficult to decide by using precision and recall, F1 Score can validate the performance of results.

$$F1\ Score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (5)$$

The main aim of the sentiment analysis is to determine people's points of view and contextual polarity. The opinion can be expressed by using subjective words and sentiment terms. However, the opinion of Tweets can be expressed by using no opinion or subjective clues. It is therefore challenging to precisely identify and recognize useful terms and opinions from the huge amounts of texts and symbols. For example, how may one detect fake news and spams, how to remove duplications and incomplete data, how to identify or understand ambiguity, irony and domain-dependent context. In this paper, we investigate the public opinion as presented in Tweets by using sentiment analysis technologies for the financial sector and identify the performance of the sentiment analysis tools.

2 Research Methodology

The aim of the methodology is to extract suitable text from Twitter and detect the sentiment scores of each Tweet. Figure 1 shows the framework of the methodology. The steps are shown as below:

Step 1: Identify promising, highly influential financial Twitter accounts.
Step 2: Identify interesting duration for sampling.
Step 3: Collect sample Tweets from Twitter.
Step 4: Remove noises and extract suitable text from Tweets.
Step 5: Generate sentiment scores by using a sentiment analysis tool (SentiStrength).
Step 6: Generate sentiment scores for each Tweets and accounts.
Step 7: Generate ground truth sentiment scores for each Tweets.
Step 8: Calculate the polarity of each financial Twitter account.
Step 9: Summarise the distribution of sentiment scores and compare predicted results and actual results.
Step 10: Analyse characteristics and evaluate the performance of tool results.

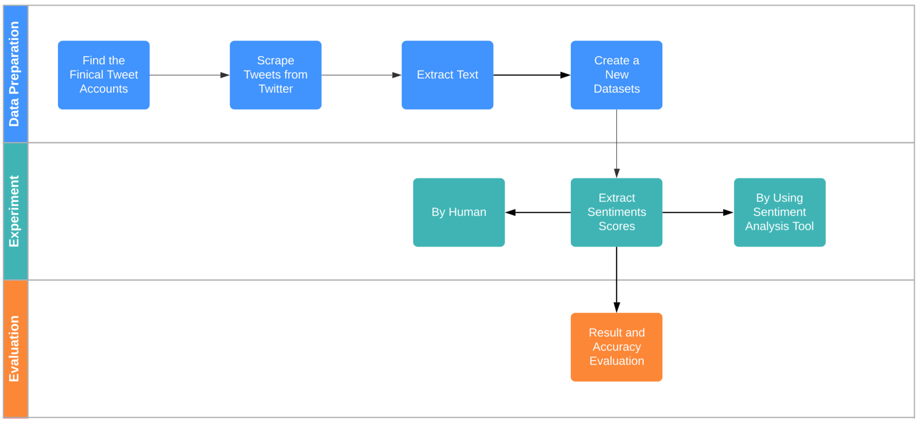

Fig. 1. Framework of Methodology

3 Experiments

3.1 Data preparation

There are several relevant Tweets datasets, such as Trump Twitter Archive [7], which includes some 50,049 Tweets from United State President Donald Trump. In addition, there are useful Python tools that scrape Tweets such as GetOldTweets3. However, most datasets are not free to use and have time restrictions. Twitter's official API also has the restriction of time that users cannot acquire older Tweets more than a week. In addition, Twitter limits each IP address to 2,000 requests per hour via their Twitter API.

In this paper, Python 3 library like GetOldTweets3 was used to scrape Tweets data from different users. It mimics a user's search using the Twitter search bar. The user scrolls down and acquires Tweets through a browser to overcome the Twitter limitations above. However, Twitter updated the API lately and these methods are no longer working. Twint, another Python tool, was used instead to overcome this limitation.

There are many influential Twitter user accounts held by individuals or financial news providers. To name a few: YahooFinance, Bloomberg, Reuters, FoxBusiness, WallStreetJournal, former US president Donald Trump, and other independent advisers and traders' accounts. For this study, five influential Twitter accounts have been selected: Bloomberg@business, @Reuters, @Forbes, @WSJ, and @realdonaldtrump. They are users who are interested in the stock market and that they may have a great influence on investors' opinions. As our Tweet samples overlap with his presidency, we particularly include the Tweets published by the former US president Donald Trump. As his Tweets, when holding such an influential role, have been observed to have significant impacts on the movements of stock market's prices. This includes Tweets indicating policy announcements ahead of their implementations.

The data set was collected from Twitter from "29th February 2020" to "3rd April 2020". This is because a significant stock market event of "trading curb" happened during this period that it is interesting to observe Tweets publications during this time. We collected 1,000 Tweets from each user to a total of 5,000 Tweets. Each Tweet includes a unique id, permalink, username, text, date, the number of reTweets, number of favorites, number of mentions, and hashtags. The texts of each Tweet are extracted and prepared for sentiment analysis.

3.2 Sentiment analysis

The main technology of a lexicon-based sentiment analysis method is to make use of a set of sentiment lexicons which includes labelled positive and negative sentiment words. The input texts are matched with these sentiment lexicons and weighted to generate their sentiment values. This represents the sentiment polarity of the text.

In this paper, the SentiStrength have been used to detect the sentiments of the Tweets. SentiStrength is a lexicon-based classifier using linguistic information to detect sentiment in textual information [5]. The results of SentiStrength are two integer scores: positive [1, 5] and negative [-1, -5]. Two scales are used because every text may include both positive and negative sentiments. It is important to detect them separately before an overall sentiment is proposed. 1 represents no sentiment and 5 strong positive sentiment. The result of 3, -5 means moderate positive sentiment and strong negative sentiment. 1 and -1 means neutral sentiment.

There are several useful features in SentiStrength. It provides a sentiment lexicon assigned with polarity and strength judgements. SentiStrength also contains booster word lists, idiom lists and negating word lists. It can identify the sentiment of common phrases and overrides individual sentiment word strengths. For example, "is like" has a score 1 (means neutral text). "like" is a comparator after "is" rather than a positive term (positive 2). The algorithm of SentiStrength detects each word to check whether there is an increase or decrease of 1. The algorithm repeats until all words are checked.

Weka is an open-source machine learning software that can be used by terminal applications, Java API, and graphical user interface. It includes a lot of built-in tools and useful packages. SentiStrength is wrapped in the AffactiveTweets Weka package and it can be accessed through the WekaPackage manager. The collections of Tweets are saved as CSV file and loaded through Weka SentiStrength Package. Next, the newly generated CSV file with sentiment scores will be analysed.

4 Results

4.1 Tweets Statistics

The outputs of SentiStrength include a positive (from 1 to 5) and a negative (from -1 to -5) polarity scores. These two scores have been aggregated to represent the sentiment composite value. Figure 2 shows the distributions of the composite and ground truth score results. Most Tweets produce a 0 (neutral) result, and that reflects the fact that most Tweets are informative and do not express sentiments directly in their texts. Ground truth values are generated based on judgements of their influence on the stock market. When a 0 value is produced, it indicates that the Tweet, based on its text and without reading into additional information such as attached links/URLs, does not have an obvious influence on stock market price movements.

The Reuters produce the most neutral scored sentiment (0) Tweets, but it is judged to have more slightly negative sentiment (-1) for its ground truth score. This indicates that although Reuters use a lot of neutral words/phrases in their Tweets, they bear a slightly negative influence on the stock market. Donald Trump uses a great number of positive words/phrases so that the positive composite sentiment scores are much higher than other users. However, when comparing them with the ground truth, their influence on the stock market is only moderately more positive.
The distributions of the composite sentiment scores among all users are very similar, except for Donald Trump. As the (former) US president, it is understandable that he may wish to project positivity into the stock market. Overall, the results are within the normal distribution, and most Tweets obtain the scores as 0. Most of them also exhibit similar distributions between the composite and ground truth scores.

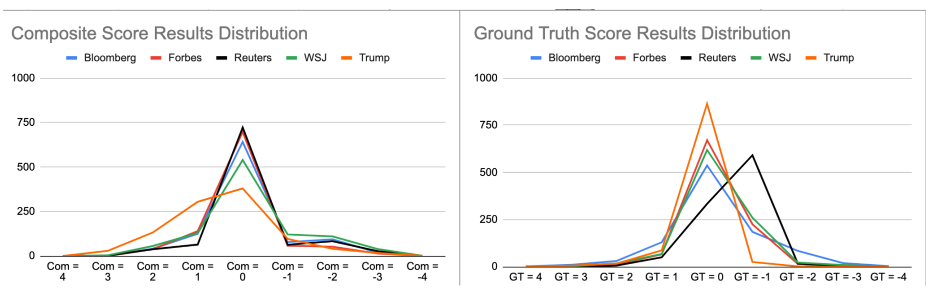

Fig. 2. Distributions of composite and ground truth score results

Table 2 gives the summary of the Sentiment Polarity Scores. Based on our samples, Donald Trump has the most positive Tweets, and WSJ has the most negative Tweets. Reuters produces the most neutral Tweets among all others (this may be an indication that it is more informative rather than judgmental). Table 3 shows the ground truth results of sentiment polarity classification. It shows that Bloomberg has the most positive scores (this indicates that although not using the most positive words/phrases, they indicate the most positive influences on the stock market). Reuters has the most negative Tweets in ground truth results. It also shows that most Twitter users have more negative sentiments towards the stock market movements, when they are compared with machine generated sentiments, except for Donald Trump.

Figure 3 shows the differences in scores between the ground truth and SentiStrength generated composite scores. There is not a big difference between them: 40.2% (difference = 0) has no disagreement and 40.4% (difference = 1) has slightly disagreement. These closely matching results may also be contributed by the fact that many of the Tweets are information based and do not contain obvious sentiments towards stock market movements or otherwise.

Table 2. Sentiment Polarity Scores

	Positive	Neutral	Negative
Bloomberg	167	642	191
Forbes	178	698	124
Reuters	104	**723**	173
WSJ	189	540	**271**
Donald Trump	**466**	380	154

Table 3. Ground Truth Sentiment Polarity Scores

	Positive	Neutral	Negative
Bloomberg	**172**	537	291
Forbes	79	671	250
Reuters	56	334	**610**
WSJ	88	619	293
Donald Trump	109	**864**	27

Degree of Disagreement

	Bloomb	Forbes	Reuters	WSJ	Trump	Total	Disagreement
Diff = 0	444	525	281	400	361	2011	No Disagreement
Diff = 1	325	343	572	354	424	2018	Slightly Disagree
Diff = 2	172	101	103	189	159	724	Partly Disagree
Diff = 3	44	26	43	51	52	216	Disagree
Diff = 4	10	5	1	4	3	23	Very Disagree
Diff = 5	4	0	0	2	1	7	Mostly Disagree
Diff = 6	1	0	0	0	0	1	Strongly Disagree
Diff = 7	0	0	0	0	0	0	Extremelly Disagree
Diff = 8	0	0	0	0	0	0	Entirelly Disagree

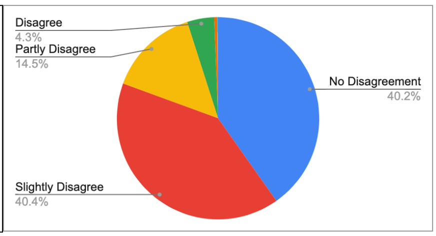

Fig. 3. Degree of Disagreement between Ground Truth and SentiStrength Scores

4.2 Accuracy and Evaluation

Table 4 presents the evaluation metrics concerning sentiment scores: accuracy, precision, recall, and F1 scores. Based on sampled Tweets, SentiStrength obtained the best performance on Donald Trump's dataset than others. The accuracy is 0.839 and it means that most sentiments have been classified correctly. This may be due to the fact that Donald Trump uses a lot of generally positive words, such as lovely, brilliantly, right, support and confidence that the sentiments of Tweets can be easier detected. Precision, recall and F1 score also confirm that most of the texts have been classified correctly.

Although comparably higher on the Recall counts, other performance indicators for Forbes are at the lowest. This may be due to the fact that most of the texts from Forbes merely states the facts, but they do not include obvious sentimental or subjective words. For example, the below Tweets from Forbes:

"The stock market bounced back today amid reports that the Trump administration is making progress on plans for a massive fiscal stimulus package that could exceed $1 trillion in an effort to reinvigorate the U.S. economy"

"Facebook has announced a $100 million grant for small businesses being impacted by COVID-19"

It is not difficult for an investor (inc. a novice one) to know that these two texts would have a very positive effect on the stock market. But the results of SentiStrength is neutral. In addition, for SentiStrength, it will be difficult to judge the polarity of the texts within a given context. It classifies sentiments based on words and phrases. Although it can identify some idioms, it does not explore the meaning of a sentence beyond the literature text. In the case of Forbes, it therefore presents a greater challenge. The performance of the tool can be summarised by considering the F1 score that is a composited measurement based on Precision and Recall. Overall, the performance of determining the sentiments of Donald Trump's Tweets gained the highest scores.

Table 4. SentiStrength Performance Measurements

	Accuracy	Precision	Recall	F1 Score
Bloomberg	0.690	0.746	0.852	0.796
Forbes	<u>0.412</u>	<u>0.387</u>	**0.869**	<u>0.5355</u>
Reuters	0.435	0.394	0.836	0.5357
WSJ	0.638	0.737	<u>0.760</u>	0.748
Donald Trump	**0.839**	**0.980**	0.852	**0.911**

Figure 4 shows the box plot diagram of the ground truth and composite scores - the differences of their median is less than 1. Most medians are between 0 to -1, except for

Donald Trump's composite score that is observably higher (0.425). It shows most texts have neural or negative impact on the stock market. It also shows that the automatically generated results from SentiStrength and ground truth exhibit similar distributions. But the accuracy depends on whether they describe texts beyond their direct meaning, and some facts would have a bigger influence than the detection of the SentiStrength.

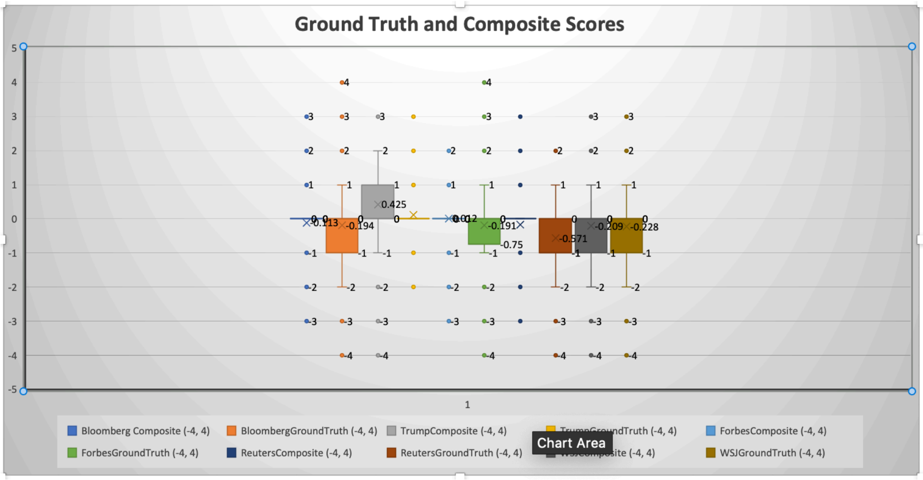

Fig. 4. Box Plot of Ground Truth and SentiStrength Scores

5 Conclusion and discussion

One of the goals of this research is to determine the influence of Tweets on the stock market, especially for news providers and well-known finance influential Twitter users. One way of doing this is to determine the sentiments of their Tweets. Twint is a Python scraping tool and was utilized to extract old Tweets from Twitter. The data were collected from five accounts: Forbes, Reuters, WSJ, and Bloomberg and the former US president Donald Trump. The datasets were used for discovering the relationships between Tweets and the stock market using sentiment analysis (SA) techniques.

SentiStrength is lexicon-based SA method and it was used to recognize and extract facts and opinion from Tweets. It assigns a positive, negative or neutral score to each Tweet. Based on the polarity of their Tweets, we determined that most Tweets of the four news providers above are either neural or slight negative. They mostly provide information without expressing obvious opinions. In comparison, Donald Trump expresses more positive opinions and they do cause positive effects on the stock market.

The performance of the SentiStrength was measured by using four metrics: accuracy, precision, recall, and F1 score. The Forbes Tweets obtained the lower accuracy than others and most of the SentiStrength scores are neutral. These results show that news providers prefer to describe facts with neutral words, and they do not express their opinion with obvious direction. However, if readers can understand the context and stock market, some neutral text (as identified by SentiStrength) can be classified as negative or positive easily. The best accuracy results of SentiStrength are obtained on Tweets of Donald Trump, because they deliver his opinion more clearly.

It is challenging to judge text precisely and acquire correct sentimental polarity from neutral words. Many Tweets may have meanings beyond the text. SentiStrength can categorize sentiments based on words, but it is difficult to explore the concealed information beyond the actual concepts of sentences. Some sentences are also complicated, and they cannot be easily judged accurately. The lexicon of sentiments might be quite general which is not specific to finance or stock market. The domain-related sentiments may require more domain-specific words and meanings and for specific contexts. The edges of lexicon would need to be expanded and improved.

Future work will also focus on understanding beyond the text such as sarcasm and ambiguous words. The sentences should be classified based on words in context and their weighting on the whole sentence so to improve the polarity classification. In this paper, the sentiment scores of SentiStrength have a maximum 0-20% differences with ground truth. More syntactic patterns can be added to indicate subjectivity and sentiments to improve the accuracy. The lexicon can be improved to include more domain-related words, phrases, and idioms to enhance the performance of a specific domain. Furthermore, ontological based methods may be utilized to enable domain-specific sentiment annotations and to create a support system by combining knowledge from various ontological sources.

References

1. Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of computational science*, *2*(1), pp1-8, March, 2011.
2. Li, Q., & Shah, S. (2017). Learning stock market sentiment lexicon and sentiment-oriented word vector from stocktwits. Proceedings of the 21st Conference on Computational Natural Language Learning, 3-4th August, 2017.
3. Mouthami, K., Devi, K. N., & Bhaskaran, V. M. (2013). Sentiment analysis and classification based on textual reviews. 2013 international conference on Information communication and embedded systems, 21-22nd February, 2013.
4. Pagolu, V. S., Reddy, K. N., Panda, G., & Majhi, B. (2016). Sentiment analysis of Twitter data for predicting stock market movements. 2016 international conference on signal processing, communication, power and embedded system, 3-4th October, 2016.
5. Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, *63*(1), 163-173, December, 2012.
6. TrumpArchive. (2021). *https://www.thetrumparchive.com/*. https://www.thetrumparchive.com/ (last accessed 7th February 2021).
7. Wen, M., Li, P., Zhang, L., & Chen, Y. (2019). Stock market trend prediction using high-order information of time series. *IEEE Access*, *7*, 28299-28308, 26th February 2019.