JOURNAL OF APPLIED INTELLIGENT SYSTEMS & INFORMATION SCIENCES
VOL. 2, ISSUE 1, PP. 22-28, JUNE 2021.
Available at: www.journal.research.fanap.com
DOI: https://10.22034/jaisis.2021.290661.1029

S&P500 INDEX DIRECTION PREDICTION USING TEXTUAL TWEETS AND THEIR CORRESPONDING SENTIMENT

Parman Mohammadalizadeh [1], Mohammadjavad Jafari [2,*]

[1] *Department of Computer Science, Buali-Sina university, Hamedan, Iran*
[2] *Department of Information Technology Management, Science and Research Branch, Islamic Azad university, Tehran, Iran*

ABSTRACT

In this paper, a novel method is proposed to predict the direction of Standard & Poor 500 (S&P500) index using the tweets in this regard as well as the index amount from the day before. At the beginning, using a dataset of all tweets and their corresponding posting times about S&P500 index, companies and securities are considered as features of the study. Next, these feature vectors are assigned three different labels based on the direction of the index change from the day before and whether the change is significant enough, creating a classification problem. Building a sentiment analysis tool based on T5 transformer which attempts to combine all the downstream tasks into a text-to-text format, sentiment feature is added to each tweet in the dataset. Lastly, after balancing the data and preprocessing the textual information through an NLP pipeline, a deep neural network is proposed to classify the processed data. It is shown that using the tweets and their corresponding sentiments, the proposed method for movement direction prediction of the S&P500 index outperformed other existing models.

***KEYWORDS*:** Natural Language Processing, Predictive Modeling, Sentiment Analysis, Deep Neural Networks, Transformer Learning

1. INTRODUCTION

Investing in stocks is one of the most exciting and trending yet very risky decisions of almost any active individual in this industry. Decision making based on information at hand is the key part of any investor or trader, which may lead to tremendous profit or partial or complete loss of the individual's account balance. It comes with great importance that human error in such decision-making scenarios is reduced to maximize the profit (Asur & Huberman, 2010). On the other hand, it is quite impossible for any human to have access and be able to consider all of the important information out there in their decision-making. This causes the utilization of machine learning and predictive modelling to be one of the hottest fields of research in financial markets.

Although the random walk theory (Malkiel, 1999) proposes a hypothesis that claims stock prices are defined purely random and they are impossible to forecast, thanks to advancements in machine learning and the growth of information available to investors it is now possible to forecast the stock prices and trends predictions with a better accuracy than that of a random approach (Bollen et al., 2011; Khashei & Bijari, 2011; Nassirtoussi et al., 2015).

* *Corresponding Author, Email: mohammad.jafari@srbiau.ac.ir*

RECEIVED: 18 MAY, 2021; **ACCEPTED**: 16 JUNE, 2021; **PUBLISHED ONLINE**: 20 JUNE, 2021

Many investors and enterprises are constantly looking for machine learning-based solutions to leverage their process of investing and making bigger profits. It is important to take into consideration that the historical values and prices are not the only motives of changes in financial markets. One of the key factors that affect the market is the supply and demand. In this study, it is shown that social media information can be a good reflection of supply and demand. Financial markets activists often utilize three different approaches to make a prediction. The first approach is called the technical analysis, which is based on the premise that the future behavior of market trends is conditioned to its historical values. The second approach, which we focus on, is fundamental analysis, any political or economic factor is considered as fundamental information regarding a company. (Nofsinger, 2001) shows that in some occasions, investors tend to buy after positive news resulting in a pump in demand. It can be shown that tweets and public sentiment can also be a reflection of fundamental information that can be used to predict the market trend. Finally, the last approach is hybrid of both using the historical price data along with textual data that mostly comes from news text (Ding et al., 2015).

The goal of this research is to show how the tweets released within a day affect the direction of the S&P500 index open value in the following day. In order to make predictions, tweets pointing out to S&P500 index or it's included companies are acquired from the dataset released by (Taborda, 2021). In this paper we showed that, tweets' textual data can be a great reflection of the market sentiment and can lead to a better trend prediction in comparison to methods that only take into account the historical price data. Our model outperformed other existing models doing the same task significantly.

2. RELATED WORK

Extensive studies have been done in this field. In many of these studies historical price data and technical indicators are used as features. Some researchers have applied machine learning techniques on historical price data, volume, average variance, etc. (Chen & Bahsoon, 2013; Huang el al., 2005; Kouloumpis et al., 2011; Pak & Paroubek, 2010; Shen et al., 2012). Deep neural networks and sentiments analysis pretrained models are used to improve viability of the readings (Shynkevich et al, 2015). Evolutionary computation through genetic algorithms are used in (F. Allen and R. Karjalainen, 1999), statistical learning using support vector machines are proposed in (Kim, 2003). Textual analysis and component modeling on news data are discussed in (Melo, 2012).

There are also some deep learning based studies like (Batres-Estrada, 2015) that utilized Deep Belief Networks, which is composed of stacked Restricted Boltzmann Machines, coupled to a Multi-Layer Perceptron. Long Short Term Memory (LSTM), is yet another well attained neural net in this field of study, it was introduced by (Hochreiter, 1997) and it gains better performance by tackling the vanishing gradient issue that recurrent networks would suffer from. Most of the studies that take textual news data as one of their main features, have leveraged LSTM in a way to improve their performance. In (Ding, et al., 2014), it is shown that the performance of daily prediction is superior to that of weekly and monthly. To map the textual information to an embedding space, multiple approaches have been used.

Other representation techniques such as word-embeddings and event-embeddings have also been used in research works (Bengio et al., 2003; Ding et al., 2015; Mikolov,et al, 2013) They outperform previous usage of textual information since they can represent complex characteristics of words or events with lower-dimensional dense vectors. In (Vargas et al., 2017) , the authors have introduced an RCNN model to forecast the intraday directional movements of the Standard & Poor's 500 index. They exploit seven technical indicators extracted from the target series and combine it with financial news titles, which they claim can be enough to reflect the whole effect of the news data. Another study has introduced a model based on Generative Adversarial Networks to predict the different financial market indices like S&P500, NYSE, and Shanghai Composite Index of China (Zhang et al., 2019).

3. METHODOLOGY AND IMPLEMENTATION

A classification model was designed based on deep neural networks to perform predictions of S&P500 index directional movement. In other words, it attempts to determine if the value of S&P500 index changes more than

0.5 percent in positive or negative direction or stays neutral in this range. The model diagram is shown in Fig. 1. Three class labels were annotated to each row of the mentioned dataset by simply comparing the S&P500 index value to that of the next day. Class labels are -1, 0 and +1, which respectively indicate negative change of more than 0.5 percent, no significant change (any change between -0.5 percent to +0.5 percent) and positive change of more than 0.5 percent. These labels are then balanced to avoid bias toward any class. Additionally, to further augment the features of this dataset, a sentiment analysis was performed on the tweets using a pretrained transformer model, which is discussed in details in next section of this study. Next, since the dataset is not balanced, to make sure there is no bias toward any of the classes, we balanced the data. Then, after mapping the textual information of the tweets to another embedding space and normalization, the data is ready to be fed into a classifier. A deep neural network is proposed to serve the task of classification in this study. Lastly, the classification performance is measured with accuracy and F-measure as performance criteria.

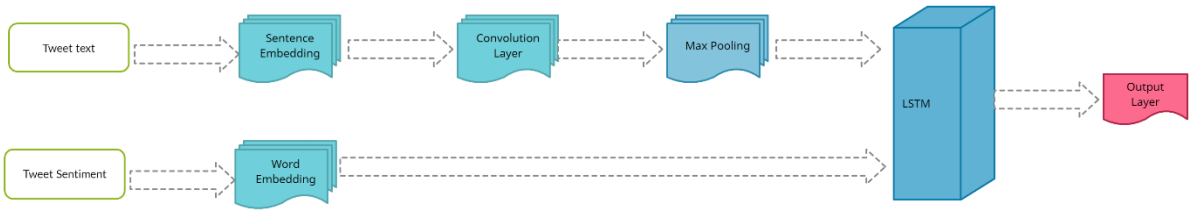

Fig. 1. Proposed model

3.1. dataset

The dataset used in this study is released by (Taborda, 2021) and consists of 943,672 tweets between April 9 and July 16, 2020, using the S&P 500 tag (#SPX500), the references to the top 25 companies in the S&P 500 index, and the Bloomberg tag (#stocks). 1,300 out of the 943,672 tweets were manually annotated in positive, neutral, or negative classes and a second independent annotator reviewed the manually annotated tweets.

The tweets retrieved were filtered out for the English language. Data collection was performed from April 9 to July 16, 2020, using the following Twitter tags as search parameter: #SPX500, #SP500, SPX500, SP500, $SPX, #stocks, $MSFT, $AAPL, $AMZN, $FB, $BBRK.B, $GOOG, $JNJ, $JPM, $V, $PG, $MA, $INTC $UNH, $BAC, $T, $HD, $XOM, $DIS, $VZ, $KO, $MRK, $CMCSA, $CVX, $PEP, $PFE. Due to the large number of data retrieved in the RAW files, it was necessary to store only each tweet's content and creation date.

3.2. Data processing

To create class labels for each row in the dataset, the S&P500 index value of the corresponding day was compared to that of the next trading day. Next, stop words which are simply frequently used words in a language that usually do not change the meaning of a sentence, are removed. In this study, we used the stop words list of SpaCy library to filter out these words in our textual data. Then, all the remaining words were stemmed and lemmatized as another level of preprocessing of our NLP pipeline. Stemming is the process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of words known as a lemma. Lemmatization collects all inflected forms of a word in order to break them down to their root dictionary form or lemma. Words are broken down into a part of speech (the categories of word types) by way of the rules of grammar. The preprocessing pipeline of this research is shown in Fig. 2. At last, using a T5 model that is by default pretrained on IMDB to be able to perform sentiment analysis task, sentiments of each tweet were added as another feature to the dataset.

3.3. Text-to-Text Transfer Transformer (T5) in details

This model was first proposed by (Raffel et al., 2019). With the thriving of Transfer Learning, Deep Learning has achieved many milestones. More specifically, in NLP, with the rise of the Transformer (Vaswani et al., 2017), various approaches for 'Language Modeling' have arisen wherein we can benefit from transfer learning

by pre-training the model for a generic task and then fine-tuning it on specific downstream problems. The model structure is a standard sort of encoder-decoder transformer.

T5 model structure is shown in Fig. 3. T5 uses common crawl web extracted text and the authors applied a simple heuristic filtering. T5 removes any line that did not end in a terminal punctuation mark. It deduplicates the dataset by taking a sliding window of three sentence chunks and deduplicated it so that only one of them appeared in the dataset.

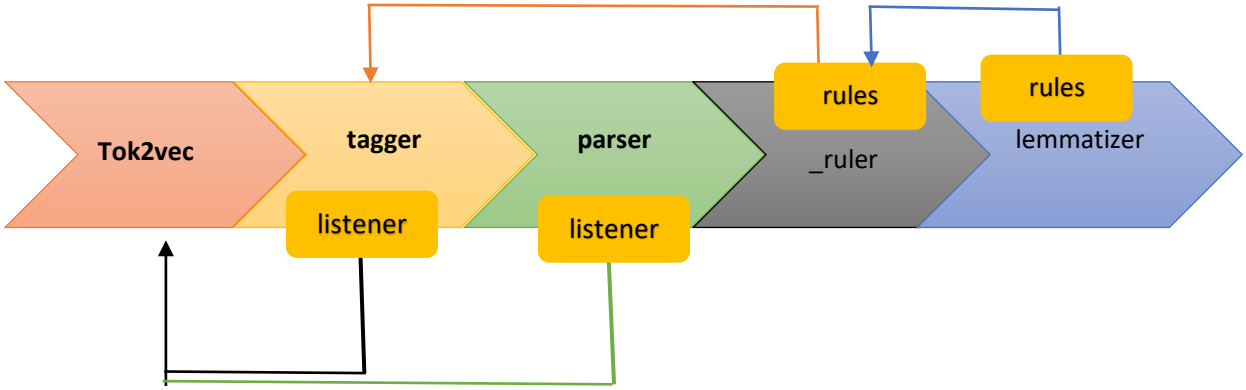

Fig. 2. NLP preprocessing using SpaCy library

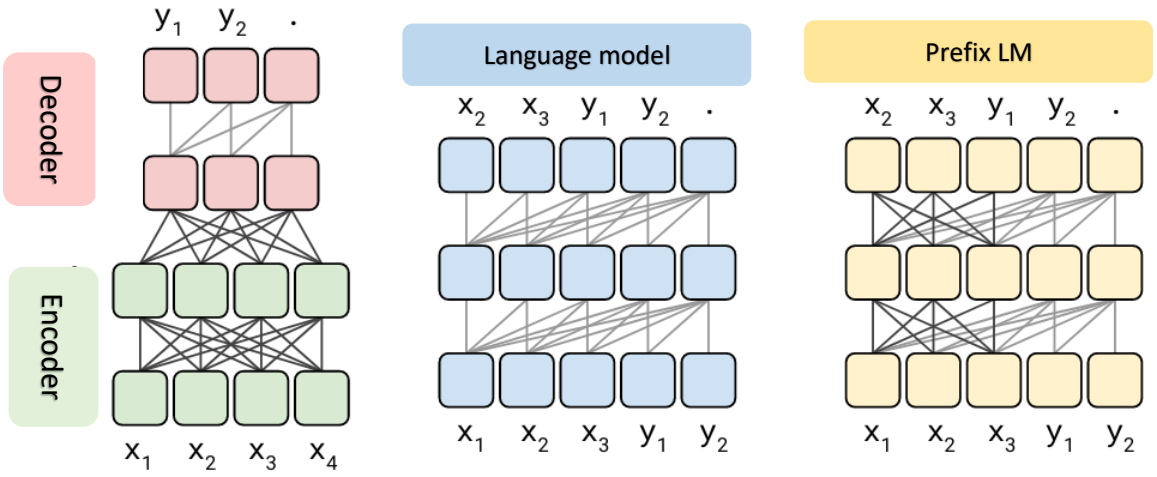

Fig. 3. T5 model structure

3.4. Classification deep neural network model

The provided model in this research is inspired and built on the base of the proposed model by (Vargas et al., 2017) and uses two types of inputs, the first one is the corresponding sentiment of each tweet which is a categorical value between [positive, neutral, negative] . The second input of the model is the textual information of the tweets, which is mapped to a sentence embedding. Given the restriction on the length of the tweet and that they usually consist of one sentence, we have used sentence embedding in this manner. In order to differentiate these two inputs, they are renamed as sentiment layer and embedding layer respectively.

The embedding layer takes a sequence of encoded sentences that corresponds to each tweet in the dataset. To obtain the encoding of the sentences, first the famous word2vec model was utilized to generate word embeddings. At last, we take an average of all the word vectors so a unique vector is obtained for each tweet. Usage of word2vec model can bring linguistic regularities benefits such as semantic and syntactic regularities.

The output of the embedding layer is then fed to a convolutional layer which is composed of four components: convolution, pooling, activation, and dropout. To capture local information through the different sentence vectors in a window, the temporal convolution that performs a one-dimensional convolution is used. The sentiment layer also goes through a word embedding layer.

Next, to be able to interpret the output of the convolution and sentiment layer as sequence of time steps, a LSTM were used. The LSTM equations are shown in equation (1) to (6), where v_t is the input of the LSTM, h_t is the output of the recurrent unit and W are the weight matrices.

$$f_t = \sigma\left(W_{vf} v_t + W_{hf} h_{t-1} + b_f\right) \quad (1)$$

$$i_t = \sigma(W_{vi} v_t + W_{hi} h_{t-1} + b_i) \quad (2)$$

$$\tilde{C}_t = \tanh(W_{vC} v_t + W_{ho} h_{t-1} + b_c) \quad (3)$$

$$o_t = \sigma(W_{vo} v_t + W_{ho} h_{t-1} + b_o) \quad (4)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (5)$$

$$h_t = o_t * \tanh(C_t) \quad (6)$$

Lastly, a fully connected layer with SoftMax activation function, whose output is the probability distribution of the class labels, is utilized to indicate the final output of the network.

4. EVALUATION

Experiments were carried out to predict the direction of S&P500 from April 2020 to July 2020. The dataset was partitioned into a training set of three months and a test set of one month. However, we only use textual data in this study, but our performance is compared to the studies that predicted the direction of S&P500 index even using different inputs and time intervals. To further validate the model, a 5-fold cross validation was performed in 10 epochs whose results are shown in Fig. 4 and Fig. 5. The model obtained an accuracy of 70.26 and F-measure of 48.10. A computer with the following specification is used to run the tests: Suprim X 3070 RTX, 64 GB RAM 4400 MHz. The experimental results are show in Table 1.

Market prediction is important for investors, traders and financial enterprises and policymakers. This work has presented a deep learning model for S&P500 index direction prediction using solely tweets reflecting the relative market sentiment. Adding the sentiment to each record of the dataset using novel text-to-text transfer transformer model is one of the main contributions of this work. The results of our proposed model shows that it performs better than other similar models of the literature. This result further approves the hypothesis that the sentiment expressed in the social media can be a good reflection of market sentiment and has temporal effect on the market. In general, the model successfully achieved the objective of this research to predict the direction of the S&P500 index.

Table 1. Results of S&P500 index prediction

Model	Accuracy
BW-SVM (Luss & d'Aspremont, 2015)	56.38%
E-NN (Ding et al., 2014)	58.83%
WB-NN (Ding et al., 2015)	60.25%
WB-CNN (Ding et al., 2015)	60.57%
E-CNN (Ding et al., 2015)	61.45%
EB-NN (Ding et al., 2015)	62.84%
EB-CNN (Ding et al., 2015)	64.21%
SI-RCNN (Vargas et al., 2017)	62.03%
WI-RCNN (Vargas et al., 2017)	61.31%
(Kamalov et al., 2020)	55.00%
Our model	**70.26%**

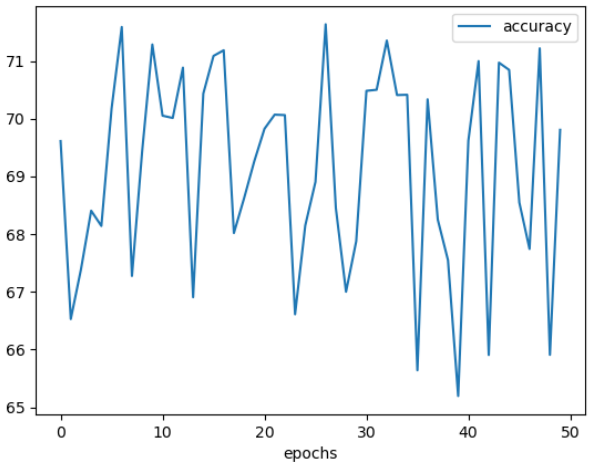

Fig. 4. Accuracy of the model in cross validation phase

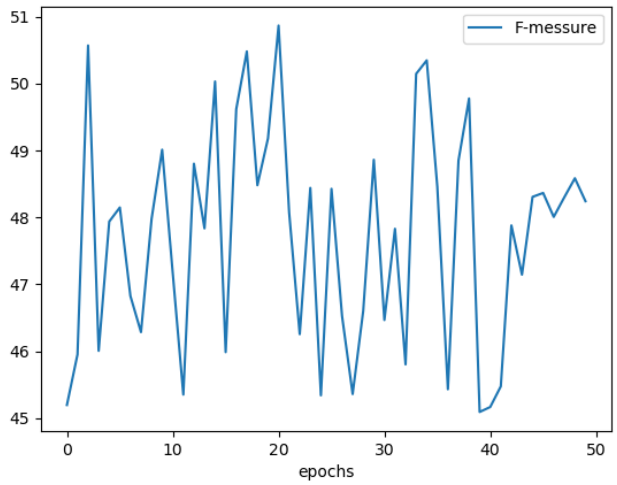

Fig. 5. F-measure of the model in cross validation phase

As a future objective, this model can consider specific stock tickers to provide better decision support for investors. We can also customize the proposed approach for different markets and time intervals. Further, to evaluate this work in a more realistic way, it can be implemented into a simple trading strategy to report the real yields and return on investments. The key to get better predictions lies in adding new textual data from sources other than twitter that can also help generalize the proposed method.

REFERENCES

Asur, S., & Huberman, B. A. (2010, August). Predicting the future with social media. *In 2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology*, (Vol. 1, pp. 492-499). IEEE.

Batres-Estrada, B. (2015). *Deep learning for multivariate financial time series*.

Bengio, Y., Ducharme, R., Vincent, P., & Janvin, C. (2003). A neural probabilistic language model. *The journal of machine learning research*, 3, 1137-1155.

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

Bruno Taborda, Ana de Almeida, José Carlos Dias, Fernando Batista, Ricardo Ribeiro. (2021). Stock Market Tweets Data. *IEEE Dataport*. https://dx.doi.org/10.21227/g8vy-5w61

Chen, T., & Bahsoon, R. (2013, May). Self-adaptive and sensitivity-aware QoS modeling for the cloud. *In 2013 8th International Symposium on Software Engineering for Adaptive and Self-Managing Systems* (SEAMS) (pp. 43-52). IEEE.

Ding, X., Zhang, Y., Liu, T., & Duan, J. (2014, October). Using structured events to predict stock price movement: An empirical investigation. *In Proceedings of the 2014 conference on empirical methods in natural language processing* (EMNLP) (pp. 1415-1425).

Ding, X., Zhang, Y., Liu, T., & Duan, J. (2015, June). Deep learning for event-driven stock prediction. *In Twenty-fourth international joint conference on artificial intelligence*.

F. Allen and R. Karjalainen. (1999). Using genetic algorithms to find technical trading rules. *Journal of Financial Economics*, vol. 51, no. 2, pp. 245 – 271, 1999.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780.

Huang, W., Nakamori, Y., & Wang, S.-Y. (2005). Forecasting stock market movement direction with support vector machine. *Computers & operations research*, 32(10), 2513-2522.

Kamalov, F., Smail, L., & Gurrib, I. (2020, December). Forecasting with Deep Learning: S&P 500 index. *In 2020 13th International Symposium on Computational Intelligence and Design* (ISCID) (pp. 422-425). IEEE.

Khashei, M., & Bijari, M. (2011). A novel hybridization of artificial neural networks and ARIMA models for time series forecasting. *Applied Soft Computing*, 11(2), 2664-2675.

Kim, K.-j. (2003). Financial time series forecasting using support vector machines. *Neurocomputing*, 55(1-2), 307-319.

Kouloumpis, E., Wilson, T., & Moore, J. (2011, July). Twitter sentiment analysis: The good the bad and the omg!. *In Fifth International AAAI conference on weblogs and social media*.

Luss, R., & d'Aspremont, A. (2015). Predicting abnormal returns from news using text classification. *Quantitative Finance*, 15(6), 999-1012.

Malkiel, B. G. (1999). A random walk down Wall Street: including a life-cycle guide to personal investing. *WW Norton & Company*.

Melo, B. (2012). Considerações cognitivas nas técnicas de previsão no mercado financeiro. *Universidade Estadual de Campinas*.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. arXiv preprint arXiv:1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *In Advances in neural information processing systems* (pp. 3111-3119).

Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2015). Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment. *Expert Systems with Applications*, 42(1), 306-324.

Nofsinger, J. R. (2001). The impact of public information on investors. *Journal of Banking & Finance*, 25(7), 1339-1366.

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).

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:1910.10683.

Shen, S., Jiang, H., & Zhang, T. (2012). Stock market forecasting using machine learning algorithms. Department of Electrical Engineering, Stanford University, Stanford, CA, 1-5.

Shynkevich, Y., McGinnity, T. M., Coleman, S., & Belatreche, A. (2015, December). Predicting stock price movements based on different categories of news articles. *In 2015 IEEE Symposium Series on Computational Intelligence* (pp. 703-710). IEEE.

Vargas, M. R., De Lima, B. S., & Evsukoff, A. G. (2017, June). Deep learning for stock market prediction from financial news articles. *In 2017 IEEE international conference on computational intelligence and virtual environments for measurement systems and applications* (CIVEMSA) (pp. 60-65). IEEE.

Zhang, K., Zhong, G., Dong, J., Wang, S., & Wang, Y. (2019). Stock market prediction based on generative adversarial network. *Procedia computer science*, 147, 400-406.