ISSN 2347 - 3983
Volume 8. No. 9, September 2020
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter85892020.pdf
https://doi.org/10.30534/ijeter/2020/85892020

Stock Price Prediction Using BERT and Word2Vec Sentiment Analysis

Willem Sebastian[1], Sani M. Isa[2]
[1]Computer Science Department, Binus Graduate Program – Master of Computer Science, Bina Nusantara University, Jakarta Indonesia 11480, willem.sebastian@binus.ac.id
[2]Computer Science Department, Binus Graduate Program – Master of Computer Science, Bina Nusantara University, Jakarta Indonesia 11480,sani.m.isa@binus.ac.id

ABSTRACT

This paper experiments with machine learning algorithm and twitter sentiment analysis to predict future stock market prices. The prediction of stock market always considered as challenging task in financial time series prediction given how dynamic stock markets are. Stock markets are heavily sentiment driven and along with rapid evolution in NLP, helping researcher in understand market sentiment better so it can give better result in stock prediction. Word2vec introduced in 2013 is touted as one of the biggest breakthrough in NLP field. However, in 2018 BERT was introduced and achieved state of the art result. The aim of this paper is to compare those method and apply those method in predicting future stock price.

Key words: BERT, Word2Vec, Stock, Sentiment Analysis, SVR, Supervised Learning.

1. INTRODUCTION

There is one thing that cannot be separated from stock which is profit. Profit is one of top reasons people invest in stock. However it is not an easy task to get profit consistently. Lot of study has been conducted to find best models to predict stock prices. Given the attractiveness of the research area, the number of successful research paper is still quite low [10]. Investors and researchers have always found forecasting trend or future value of stocks an interesting topic.

Stock market are known for having high risk high return characteristic. There are lot of factors that affect stock price. One of them is people sentiment. According to [1], sentiment posted on social media has a correlation with how market as a whole will move. Basically, if people optimistic and give a lot of positive tweet, then value of stock will go up and vice versa.

Nowadays, people can easily post their thoughts on social media. Twitter is one of most notable social networking and connect with each other real time with more than million active users. This makes twitter like huge library that contains massive information or data inside. It is like gold mine for researcher in NLP.

Sentiment Analysis or opinion mining getting popular in recent years. As social media platform grow bigger, mineable text also increased. Sentiment analysis can determine people sentiments and emotions. With rapid development in Natural Language Processing, in 2013, Tomas Mikolov introduced Word2Vec [2], technique for NLP that uses a neural network model to learn word associations from a large corpus of text. This model can detect similar words or suggest additional words for a partial sentence. This Word2Vec quickly become state of the art in NLP. However, in 2018 Google introduce new word embedding technique called Bidirectional Embedded Representational from Transformer (BERT)[4] that achieve state of the art result on eleven natural language processing.

In this paper, we compare BERT and Word2Vec model to predict future stock price. We apply Word2Vec and BERT to sentiment classification. Our purpose is to find is BERT better than Word2Vec in sentiment classification and can BERT give better result in future stock price prediction than Word2Vec.

2. RELATED WORKS

Forecasting stock has attracted many researcher in recent years. Rapid development on NLP has inspired researcher to maximize text into information.[9, 14, 16, 17, 19] use news sentiment analysis to predict future stock. [1, 5, 15, 19] use twitter micro blogging platform to predict future stocks. While [6, 12, 13] utilized StockTwits through pipeline API, processed them for NLP and sentiment analysis.

[7] Proposed approach with Word2Vec model and used model to predict trading actions. Their results shows Word2Vec achieved positive results in all scenarios while the average yields of Moving Average and MACD still lower compared to Word2Vec.

[8] Tried to combine BERT and Long Short Term Memory (LSTM) to predict three stocks that are listed in Hongkong Stock Exchange which are Tencent, CCB, and Ping an. Their results shows BERT give better performance compared to other models, such as FastText and Transformer + attention.

[11] Tried to predict direction of stocks in Turkish stock market. They compared model performance between BERT, LSTM, RNN, and CNN. Their results show BERT is the best performing model when average accuracy results of each model are considered with 96.26% accuracy value.

3. PROPOSED METHOD

In this section, we describe the proposed methodology for predicting future stock price through sentiment analysis. The proposed methodology is carried out in four steps(figure 1).

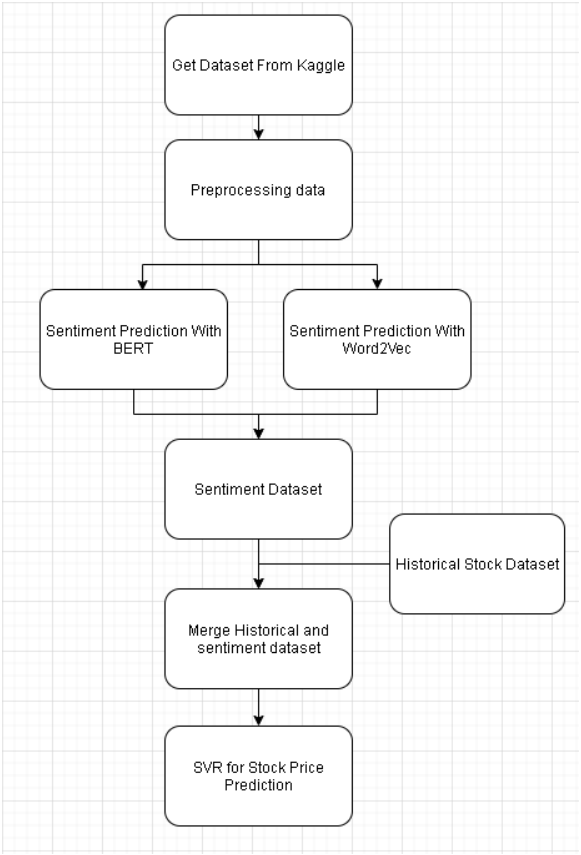

Figure 1: Flowchart of proposed Methodology

3.1Get Dataset from Kaggle

In this study, we use two data. One is historical data taken from Yahoo finance API, and tweet data from Kaggle website. Kaggle is the world largest data science community with powerful tools and resources. Dataset we choose is taken from https://www.kaggle.com/ryanchan911/selective-stock-headlines-sentiment.It has 9344 tweet start from 10 October 2018 until18 January 2020 and separated into 5375 positive tweets and 3969 negative tweets.From dataset, we exclude tweet with Apple, Microsoft, and Nike tag. Then we use those tweet as test dataset and the rest as training dataset.

3.2. Pre-processing Data

We make different approach in pre-processing tweet with BERT and Word2Vec. Pre-processing in BERT is much shorter because BERT is trained with full sentence. In BERT pre-processing we just delete @name and trailing whitespace.
In Word2Vec pre-processing, we did several pre-processing step such as case folding, change "'t" and "not" to cannot, delete @name, delete stopwords except "not" and "can" and delete trailing whitespace.

3.3Sentiment Prediction

After pre-processing tweet data, we train our BERT and Word2Vec model.

3.3.1 Sentiment Prediction with BERT

In this step, we used BERT pre-trained from transformers library. First we used BERT tokenizer to change tweet into BERT token. Then we create BERT model using Bert Sequence Classification imported from transformers library. We trained model with 4 epochs. After model was trained using training dataset, we use test dataset which contain Apple, Microsoft and Nike stocks to our model and we get BERT sentiment dataset.

3.3.2 Sentiment Prediction with Word2Vec

In this step, we used Word2Vec imported from Gensim library. Then we train with 10 epochs. Then we use test dataset to test our model and got Word2Vec sentiment dataset.

3.4. Merge Historical and Sentiment Dataset

After we get sentiment dataset, we merged historical data and sentiment dataset from BERT and Word2Vec based on tweet matching date. However, there could be more than one tweet on each day for each company. So the data is aggregated day-wise. It means, sentiment score on that day will be average of those tweet sentiment score on same day.
In this study, we want to predict stock price one, three and seven days to the future. So we add price on day of tweet posted, price + 1 day, price + 3 days, and price + 7 days. However when tweet posted, it possible to posted on holiday. So we decide price will be price on the day before tweet posted when stock market open. It also possible to price + 1, price + 3 and price + 7 is holiday, so we decide price will be price after the day when stock market open. At the end, the file contains three attributes i.e. date, stock price, sentiment on particular day and stock price + n days which 1, 3 and 7 days respectively.

3.5 Future Stock Price Prediction with SVR

We used SVR model to predict future stock price. For the training data, we will use merged data from step 4. We conduct multiple experiment, we create multiple model trained based on Apple, Microsoft and Nike merged data. Then we trained model using "RBF" kernel.

4. RESULT AND DISCUSSION

The BERT model and Word2Vec model we build gives results as shown on table 1.

Table 1: Sentiment Classification Results

Stock	BERT Accuracy	Word2Vec Accuracy
Apple	95.20%	73.06%
Microsoft	98.46%	76.54%
Nike	95.60%	73.63%

Table 1 shows that BERT achieved outstanding result in sentiment classification task. BERT achieved better result because it trained on full sentence so it can understand context of sentence better, unlike Word2Vec which trained per word in sentence.

Table 2: MSE with SVR RBF

	Stock	MSE P+1	MSE P+3	MSE P+7
BERT	Apple	25.48	56.01	92.03
	Microsoft	0.054	3.889	8.822
	Nike	0.145	1.437	3.259
Word2Vec	Apple	21.77	53.964	94.337
	Microsoft	0.901	4.825	10.06
	Nike	0.208	1.595	3.207

From Table 2, we can conclude BERT can predict stock price better than Word2Vec. BERT can achieved lower Mean Square Error (MSE) because BERT perform better in classify tweet sentiment. Also it harder to predict in longer period of time as shown in table 2, the longer prediction time, MSE will increase.

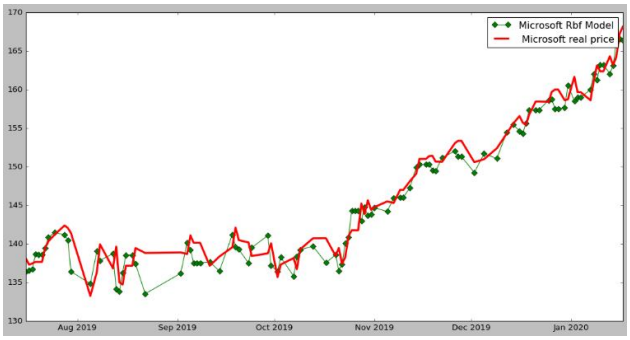

Figure 2: Microsoft stock price1 day prediction

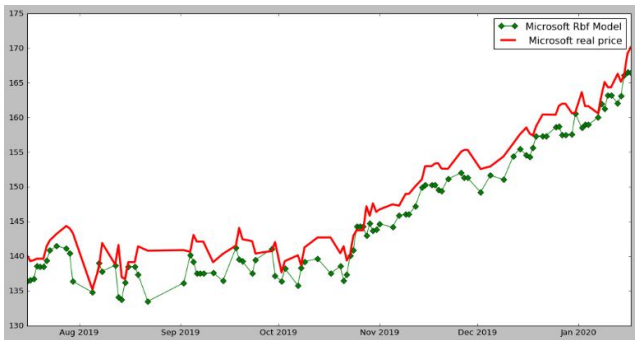

Figure 3: Microsoft stock price 3 day prediction

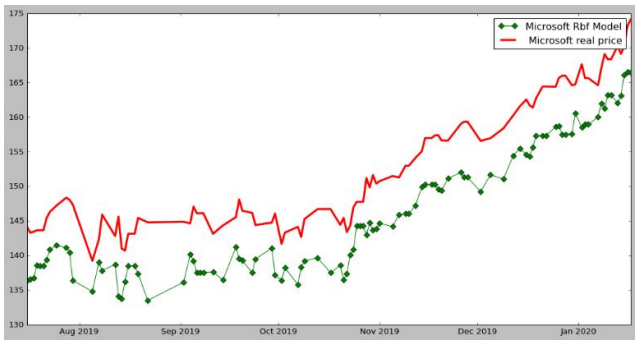

Figure 4: Microsoft stock price 7 day prediction

5. CONCLUSION AND FUTURE WORK

In this paper, we shown BERT outperform Word2Vec in sentiment classification performance. This outstanding accuracy also help in stock price prediction as it gives lower MSE. This study proves BERT as current state of the art for NLP. However, in this work, we used dataset downloaded from Kaggle website. The dataset only has 9344 tweet. The current study can be extended by using larger dataset. With increasing size of training dataset, the models tend to perform better.

REFERENCES

1. T. Sun, J.Wang, P. Zhang, Y. Cao, B. Liu and D. Wang, **Predicting Stock Price Returns using Microblog Sentiment for Chinese Stock Market.** *2017 3rd International Conference on Big DataComputing and Communications (BIGCOM), Chengdu, 2017*, pp. 87-96, August 2017.
2. M. Tomas, C. Greg, C. Kai and D. Jeffrey.**Efficient Estimation of Word Representations in Vector Space**. *Proceedings of Workshop at ICLR. arXiv:1301.3781v1*, September 2013.
3. M. Tomas, S. Ilya, C. Kai, C. Greg and D. Jeffrey. **Distributed Representations of Words and Phrases and their Compositionality**. *Advances in Neural Information Processing Systems*, pp. 3111-3119, October 2013.
4. J. Devlin,M. Chang, K. Lee and K. Toutanova.**BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.***arXiv preprint arXiv:1810.04805,* October 2018.
5. R. Pimprikar, S.Ramachandran & K. Senthilkumar. **Use of Machine Learning Algorithms and Twitter Sentiment Analysis for Stock Market Prediction**. *International Journal of Pure and Applied Mathematics, Volume 115 no. 6 2017*, pp. 521-526, 2017.
6. B. Rakhi and D. Sher. **Integrating StockTwits with sentiment analysis for better prediction of stock price movement.***2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, March 2018.
7. H. Joshua, H., Xin and M. Hao and L. Duan and W. Qi and X. Yabo. **BERT-based Financial Sentiment Index and**

LSTM-based Stock Return Predictability.*Arxiv:1906.09024*, June 2019.

8. S. Zengcai, X. Hua, Z. Dongwen and X. Yunfeng. **Chinese sentiment classification using a neural network tool — Word2vec.***2014 International Conference on Multisensor Fusion and Information Integration for Intelligent Systems(MFI),* September 2014.
9. A. Giuseppe, C. Luca, G. Paolo and B. Elena. **Combining News Sentiment and Technical Analysis to Predict Stock Trend Reversal.***2019 International Conference on Data Mining Workshops (ICDMW),* November 2019.
10. P. Marko and L. Dejan**. Discovering Language of Stocks**.*Frontiers in Artificial Intelligence and Applications-Databases and Information Systems X, 315, pp. 243 -258,* February 2019.
11. K. Zeynep, O. Derya and U. Mitat. **Financial Sentiment Analysis for Predicting Direction of Stocks using Bidirectional Encoder Representations from Transformers (BERT) and Deep Learning Models**. *International Conference on Innovative Technologies (ICIIT-19),* December 2019.
12. C. Scott, M. Praveen and C. Joseph. **Forecasting Stock Prices Using Social Media Analysis.***2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress,* November 2017.
13. B. Rakhi and D. Sher. **Integrating StockTwits with sentiment analysis for better prediction of stock price movement.** 2018 *International Conference on Computing Mathematics and Engineering Technologies* (iCoMET), March 2018.
14. H. Ziniu, L. Weiqing, B. Jiang, L. Xuanzhe and L. Tie-Yan**. Listening to Chaotic Whispers: A Deep Learning Framework for News-oriented Stock Trend Prediction.***WSDM 2018*, 261-269, December 2018.
15. P. Sasank, C. Kamal, P. Ganapati and M. Babita. **Sentiment Analysis of Twitter Data for Predicting Stock Market Movements.**2016 *International Conference on Signal Processing, Communication, Power and Embedded Systems (SCOPES),* October 2016.
16. J, Kalyani, N, Bharathi and Rao, Jyothi. **Stock Trend Prediction Using News Sentiment Analysis. International Journal of Computer Science and Information Technology.***International Journal of Computer Science & Information Technology (IJCSIT*) Vol 8, No. 3, June 2016.
17. S. Mohan, S.Mullapudi, S. Sammeta, P. Vijayvergia and D. C. Anastasiu,**Stock Price Prediction Using News Sentiment Analysis**. *2019 IEEE fifth International Conferenceon Big Data Computing Service and Applications (BigDataService)*, Newark, CA, USA, 2019, pp. 205 -208, April 2019.
18. R. Andrzej and S. Michał**. Towards Predicting Stock Price Moves with Aid of Sentiment Analysis of Twitter Social Network Data and Big Data Processing Environment**. *Advances in Business ICT: New Ideas from Ongoing Research* (pp.105-123), November 2017.
19. V. Bruce, G. Adrian, H. Geoff. Do news and sentiment play a role in stock price prediction? *Applied Intelligence,* April 2019.