2020 2nd International Conference on Economic Management and Model Engineering (ICEMME) | 978-1-7281-9144-7/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICEMME51517.2020.00206

Stock Price Prediction Based on LSTM Neural Network: the Effectiveness of News Sentiment Analysis

Yuqiao Guo
University of Illinois at Urbana-Champaign
Champaign, IL, US, 61820

***Abstract* -This paper retrieves news articles from the New York Times and conducts sentiment analysis for news headline and text body, then combine quantitative sentiment score with stock historical stock basic features together, using LSTM neural network to predict both future stock close price and stock return. The main purpose is to compare the prediction result of model which includes sentiment factors with that only considers historical stock basic features. LSTM neural network shows a good ability in long-term prediction, the experiment is based on LSTM model for three representative large companies in the US. The final results confirm that the prediction accuracy is higher for model that consider sentiment influence from website news article.**

Keywords-sentiment analysis; stock market prediction; LSTM, machine learning; NLP

I. INTRODUCTION

Stock market has a significant influence on both individual investor's interest and the country's entire economy trend. Therefore, stock movement prediction plays an important role in reducing risk of loss and increasing gains. However, stock prediction is always regarded as a challenging work. Due to the high volatility and uncertainty of complex market dynamics, internal and external information changes constantly and it is hard to make the best prediction.

In the research of stock market, one hypothesis has huge influence: Efficient Market Hypothesis (EMH) [1], which states that the security current price reflects all market information and if there is a new change in financial market, the stock price will adjust immediately. Even EMH seems to discourage the activity of stock price prediction, some researchers propose two approaches that works on stock market forecasting: fundamental and technical analysis [2]. Fundament analysis focus on making prediction from current internal and financial aspects of company, like earnings and revenue, whereas the technical analysis believes that past trend is essential and future trend relies more on historical stock activity. The existing study also proposed that the third type of analysis, social sentiment analysis is highly related to stock price movement, when public sentiment is positive, investors tend to have high and optimistic expectations to some risky assets like stock, which will cause increase in stock price and vice versa [3]. The recent global outbreak of coronavirus (COVID-19) pandemic caused widespread influence on stock markets and other financial sectors, which also shows the power of public sentiment [4]. Most previous work only consider using technical, fundamental analysis or sentiment analysis alone, this paper combines technical and sentiment analysis together. By considering both stock historical data and public sentiment based on company performance, this paper tries to find whether it could get a better result for stock price prediction.

Reflected in public emotion, public sentiment has crucial and direct impact on stock market movement and therefore it is necessary to gather those unstructured data and conduct sentiment analysis. Most current studies about the effect of sentiment on stock prediction focus on social networking service, like Twitter, because they think it provides more comprehensive information. However, the information collected from Twitter are somewhat messy and may contain lots of noise [5]. This paper chooses the New York Times as major source of text information and analyze news headlines and text body by applying Natural Language Processing (NLP) tools and finally quantify the unstructured text data.

Apart from features used to make prediction, selecting an appropriate technical predicator is also significant in the success of prediction. The recurrent neural network (RNN) is one of the most popular branches of deep learning, but due to the limited ability to memory in long-term, one of its special case, long short-term memory (LSTM) neural networks, are more suitable in this case since stock price is a long-time sequence with random and nonstationary characteristics.

This paper will investigate in whether the sentiment from company related articles from news website have predictive capabilities in stock prediction, especially for one day interval. The prediction will base on the LSTM (Long Short-term Memory) model and the overall efficacy of model will be evaluated. The final result shows that LSTM model works well in stock prediction and considering sentiment in stock prediction do improve the overall accuracy.

II. RELATED WORK AND TECHNOLOGY BACKGROUND

A. Related work

According to Fama (1970), security markets could reflect all information regards to individual stocks the whole stock market efficiently, therefore neither fundamental nor technical analysis could help investors earn more money from stock prediction. Nonetheless, many published studies strongly against this hypothesis by proposing that both technical analysis, such as historical stock price, and fundamental analysis, such as public emotion analysis, could help investors predict stock market better. From the experiment conducted by Roberto et al (2015),

978-1-7281-9144-7/20/$31.00 ©2020 IEEE
DOI 10.1109/ICEMME51517.2020.00206

Authorized licensed use limited to: California State University Fresno. Downloaded on July 01,2021 at 09:46:40 UTC from IEEE Xplore. Restrictions apply.

stock returns from the proposed trading rule based on technical analysis for the European are higher than that for the US index, which indicates the inefficiency of the stock markets. Matthias (2011) rejects the EMH based on the finding that there exists positive correlation between negative media sentiment and the decline in stock returns [6]. The study of Ranco et al (2015) shows significant evidence of correlation between stock price returns and sentiment in company related tweets from Twitter [7]. While EMH may be reasonable in some circumstance, it is still necessary for investors to construct different models to predict stock price in order to gain more return.

Given enough information as inputs, many studies in the past developed various machine learning techniques to realize stock prediction. Nikou et al (2019) compared different machine learning algorithms including Support Vector Machine, Random Forest, Deep leaning, etc. [8]. Their final result indicates that LSTM model, which is a special case of RNN that belongs to deep learning, performs better in stock close price prediction than other methods. In the research of Achkar etal (2018), they compare two kinds of neural network, Multilayer perceptron (MLP) and LSTM and the result shows that percentage error in LSTM is smaller.

B. *Long Short-Term Memory (LSTM)*

Compared to traditional neural network, recurrent neural network (RNN) has internal memory, they memorize all information stored in the past and use it to make decisions in further step. Even RNNs work well when dealing with short sequence data, they suffer from two major problems: gradient vanishing and gradient exploding.

To solve the problems of RNN, Long Short-Term Memory (LSTM) could be considered. LSTM works as a specific RNN with the improvement on identifying long-term dependency in sequence data. Different from RNNs, LSTMs perform a more complex way in computing the hidden state with replacing the traditional hidden layer neurons by sets of memory cell. They are able to regulate information by three gate structures and these gates update information selectively by learning what kind of information in the sequence is important to keep or identifying information that not useful and then throw them away.

C. *Natural language processing (NLP)*

Natural Language Processing (NLP) is a study and technology field that investigates how computers could be used to interpret human language text. As a sub-field of NLP, sentiment analysis helps people to extract emotion factor in given text. Through measuring and analyzing attitudes within the text, it provides valuable insights about the possible relation between stock movement and human sentiment, therefore enable us to make a better prediction of future stock price [9].

III. METHODOLOGY AND PREDICTION MODEL

This section will discuss the methodology in this paper. To identify the effect of social sentiment on stock prediction, this paper intends to build a forecasting model and conduct result analysis. There are five stages:

- Collecting stock historical data and news data
- Making sentiment analysis and quantify sentiment within news article
- Combining two types of data together and processing data into usable format
- Model establishment, parameters setting and model training
- Testing results comparison and prediction effect evaluation

A. *Model Selection*

Stock market is full of uncertainty and influenced by various factors, like macroeconomic policy, the performance of company and public sentiment. Due to the characteristics of high-volatile and noise, it is difficult to apply some traditional statistical models into financial time series prediction with high precision [10]. Some studies show that some machine learning method, like recurrent neural networks, could be successfully used to forecast time series [11]. As one specific case of RNN model, LSTM has unique memory selection mechanism. It stores useful memory and discard unsuitable information by using its memory cells, which could effectively capture the structure of data over long period with high accuracy. Therefore, this paper considers LSTM as a good choice to predict stock price.

B. *Data collection*

1) *Stock acquisition*

The historical daily stock data in this paper are obtained from Yahoo Finance API in Python 3.8, the basic features include the opening and closing price, the highest and lowest price among the day, and trading volume. All companies stock data are between January 1st, 2015 and August 13th, 2020.

2) *Text data acquisition*

In order to quantify the public sentiment towards company stock, this paper collect all news articles from 2015.1.1 to 2020.8.13 that contain the company name from The New York Times. In most recent studies about sentiment analysis in stock prediction, researchers regard Twitter as a valuable source that could reflect public mood. They proposed that because there is an increasing amount of people are use twitter to express their opinions for its easy accessibility, data extracted from microblogging platform is efficient for studies about marketing trend and social behavior [12]. This article, however, choose The New York Times as the text data source. Because compared to twitter, news and articles posted on The New York Times are manually screened and it could be more authoritative and accurate, which filters unrelated noise information. Due to the high amount of data, all text information is obtained by using New York Times API and web scraping. The article search API was queried for published articles discussing the selected company, each piece of news contains the information of Date, web URL and snippet. Then all articles main body were retrieved through web scrape by using each web URL from API.

C. *Sentiment analysis*

After collecting both news headlines and body content, natural language processing technology was applied with two sentiment analysis toolsets to quantify the sentiment in numeric

Authorized licensed use limited to: California State University Fresno. Downloaded on July 01,2021 at 09:46:40 UTC from IEEE Xplore. Restrictions apply.

values. VADER (Valence Aware Dictionary and Sentiment Reasoner) is a rule-based tool that provides effective analysis of sentiments from social media. It is based on a lexicon that matches features of sentence to sentiment scores, which could be obtained by adding all word's intensity in the sentence together. Because it was primarily designed to evaluate relatively short sentence, it was used to analyze the headline of each article. The corresponding package of VADER is NLTK package in Python. The output of VADER includes four values: positive, neutral, negative and compound. This paper only retains compound score because it is a metric that calculates the sum of all the lexicon ratings, and it was normalized between -1(most extreme negative) and +1 (most extreme positive). This score gives us an overall understanding of the extent of sentiment.

TextBlob is a Python library to process textual data, this paper uses it to evaluate the sentiment for each article content. This paper applies the sentiment property of this package, which returns in form of polarity score and subjectivity score. The polarity score is the value within the range [-1.0, 1.0] where 0, +1 and -1 represents neutral, very positive sentiment very negative sentiment, respectively. The subjectivity is the value within the range [0.0, 1.0] where 0 and 1 indicates very objective very subjective, respectively. Since the main purpose of this paper is to measure how sentiment could help people predict stock price better, only consider polarity score is considered here. The original output of both headline and article content sentiment analysis is shown as Table.1.

TABLE I. OUTPUT OF SENTIMENT ANALYSIS

date	TextBlob article polarity	TextBlob article subjectivity	article	compound	negative	neutral	positive	sentiment	snippet	url
2020-08-12	0.058147	0.426631	Excerpts from recent editorials in the United ...	0.9989	0.090	0.796	0.114	{'neg': 0.09, 'neu': 0.796, 'pos': 0.114, 'com...	Excerpts from recent editorials in the United ...	https://www.nytimes.com/aponline/2020/08/12/us...

Finally, in order to measure the sentiment for each trading day, results are grouped together by date. For single day that have more than one article records, this paper takes the mean sentiment score. The final data frame of sentiment analysis of news article extracted from the New York Times is shown in Table.2, it contains two features, compound score measures the sentiment for headline and polarity score measures the sentiment for the whole article body.

TABLE II. FINAL FORMAT OF SENTIMENT SCORE

Date	TextBlob article polarity	compound
2020-08-06	0.121371	0.720280
2020-08-07	0.070913	0.344415
2020-08-10	0.066128	0.690837
2020-08-11	0.072134	0.549500
2020-08-12	0.106250	0.403140

D. *Data processing and model construction*

1) *Scaling*

TABLE III. DATA FRAME OF MODEL INPUT (CLOSE PRICE PREDICTION)

Date	TextBlob article polarity	Snippet compound score	Open	High	Low	Close	Volume
2015-01-02	0.056533	0.998100	222.869995	223.250000	213.259995	219.309998	4764400
2015-01-05	0.000000	0.000000	214.550003	216.500000	207.160004	210.089996	5368500
2015-01-06	0.095295	0.494000	210.059998	214.199997	204.210007	211.279999	6261900
2015-01-13	0.134708	0.997700	203.320007	207.610001	200.910004	204.250000	4477300
2015-01-14	0.030974	0.554967	185.830002	195.199997	185.000000	192.690002	11551900
2020-08-10	0.008839	0.702800	1448.000000	1457.500000	1385.839966	1418.569946	7522300
2020-08-11	0.123980	0.893450	1396.000000	1420.000000	1365.000000	1374.390015	8625800

Before building the LSTM model, it is important to combine sentiment scores and stock historical data together and transform them into same scale range. Table.3 shows the data frame of combine data, the first two columns are sentiment score got from article sentiment analysis, rest are stock related indicators. It can be seen that the magnitude of stock price in 2015 and 2020 is different, and the data range of sentiment score is different from both stock price and stock volume. Using data

Authorized licensed use limited to: California State University Fresno. Downloaded on July 01,2021 at 09:46:40 UTC from IEEE Xplore. Restrictions apply.

in this format will negatively influence the prediction result, thus a normalization is important to eliminate the magnitude difference between data.

This paper chooses z-scale approach, using equation.1 to rescale the data from the original range, where μ is the mean of the training samples and s is the standard deviation of the training samples.

$$\frac{x_i - \mu}{s} \tag{1}$$

2) Dataset Split and parameters setting

All data collected have range from 2015.1.1 to 2020.8.13, this paper splits training and testing dataset follow the proportion 8:2. Thus, our final data used for training and testing are from 2015.1.1 to 2019.6.29 and 2019.6.30 to 2020.8.13, respectively. Within the training dataset, 20% are used for model validation.

During the model fitting, this paper tried different value for various parameters of the LSTM model in order to get better prediction accuracy. Epoch measures how many times the entire dataset pass through the neural network, improper epoch number would cause underfitting or overfitting. Due to the large size of dataset, it is impossible to pass whole dataset through the neural net at once, thus, dividing dataset into several sets (batch) is necessary. Batch size is the number of training examples in each single batch. After tuning parameters, the final LSTM model has a two-layer network structure with 50 hidden nodes per layer. The batch size is 128 and the number of epochs is 40.

3) Model Construction

This paper chooses three representative large cap stocks from the US stock market: Amazon (AMAZ), Microsoft (MSFT) and Tesla (TSLA). The reason to consider large cap stock is that for large company, the number of daily news data is large, and it will be easy to capture public sentiment. Fig.3 shows the whole structure of the prediction model in this paper, two sentiment scores and four stock indicators are used as inputs for training sets. Since both stock closing price and stock return are standard benchmarks that used by investors to measure daily stock performance, this paper uses those two features as the output of the LSTM model separately. The close price could be extracted from Yahoo Finance API directly, since people will have to buy at the day's open price and sell at the day's close price, this paper chooses formula (2) to calculate stock daily return. The timestep in this model is 1, which means all input features at day t are used to predict the close price at day t+1 and stock return at day t+1.

$$Return_t = \frac{Open_t - Close_t}{Open_t} \tag{2}$$

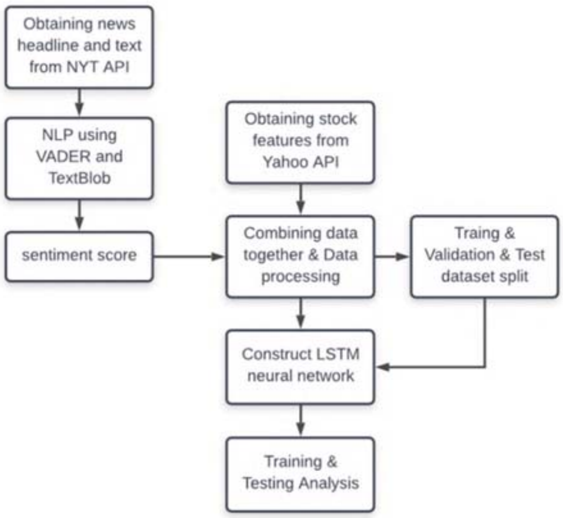

Figure 1. Whole process of model construction

IV. EXPERIMENTAL RESULT AND ANALYSIS

A. Model Training Analysis

The learning shows the learning effect of the model over time, on the train and the validation data set, a learning curve provides information to track whether the model is underfitting, overfitting or well-fitting. This paper uses training loss and the validation loss to measure model fitness, the loss value is calculated by mean squared error where value of 0 means that the dataset learned perfectly with no mistakes.

Fig. 2 show the learning curve for Tesla stock LSTM model with and without sentiment score. The x-axis is the number of experiments and y-axis is the mean square error as loss value for each experiment. From the graph, the value loss for both tarin and validation dataset are decreasing to a point and after 10 experiments, the loss is stable with value nearly 0, and there is a minute gap between two curves. This indicates that the model is stable has a good fit, which is suitable to predict stock price with high volatile.

To compare the fitness between model with and without score, it can be seen from y-axis in Fig. 2 that initial train and

Authorized licensed use limited to: California State University Fresno. Downloaded on July 01,2021 at 09:46:40 UTC from IEEE Xplore. Restrictions apply.

validation loss are bigger for model without sentiment score than the model with sentiment score. In addition, Table.4 shows that the MSE value for train and validation dataset in model with sentiment score is smaller than that without score, this comparison indicates that the training and validation effect is better in model that includes sentiment score.

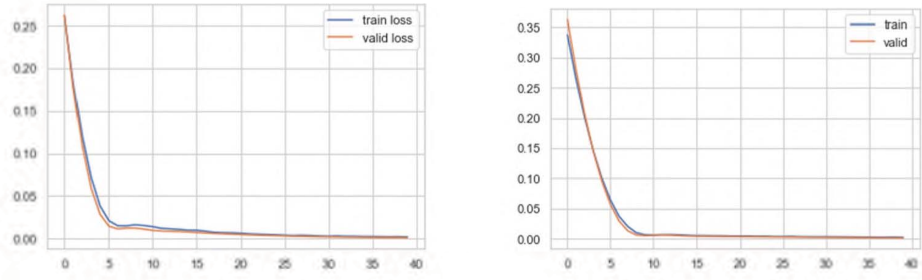

Figure 2. Train and validation loss for Tesla stock

TABLE IV. MSE VALUES FOR TRAIN AND VALIDATION DATASET

MSE Value	Train	Validation
With Score	0.002	0.0011
Without Score	0.0024	0.0015

B. Prediction result analysis

The performance of the LSTM stock price prediction model could be evaluated based on the mean square error (MSE) and mean absolute error (MAE). MSE represents the average of differences between actual and predicted values in dataset. MAE measures the average of the absolute value of differences for all predictions and doesn't consider the direction. As in the formula below, n denotes the total number of observations in the model, y_i and $\hat{y}_i$ refers to actual and predicted value, respectively. Smaller MSE and MAE indicates that predicted values are closer to true values, model with smaller MSE provides a better prediction accuracy. From Table.5 and Table.6, for each three company, both MSE and MAE values for model with sentiment score are smaller than that without sentiment score. By comparing these two results, adding sentiment score of company related news could effectively improve the prediction accuracy of stock close price and stock return. Moreover, based on the small value of MSE and MAE, the training LSTM model in this experiment could give a good prediction result.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \qquad MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Fig. 3-5 show the respective predicted closing price of Amazon, Microsoft and Tesla. The x-axis is the timestamp in testing dataset and the y-axis represents the value of stock close price on each day. The blue, orange and green lines refer to actual closing price, close price predicted by model with sentiment score and without sentiment score, respectively. It can be noticed that the predicted close price using sentiment score model for all three companies are much closer than true value, which means sentiment score do have positive effect in stock prediction and this is correspond to the result given by MSE and MAE comparison. Fig. 6 shows the prediction result for Microsoft stock return with sentiment score, which also indicates that LSTM model works well in prediction of daily return.

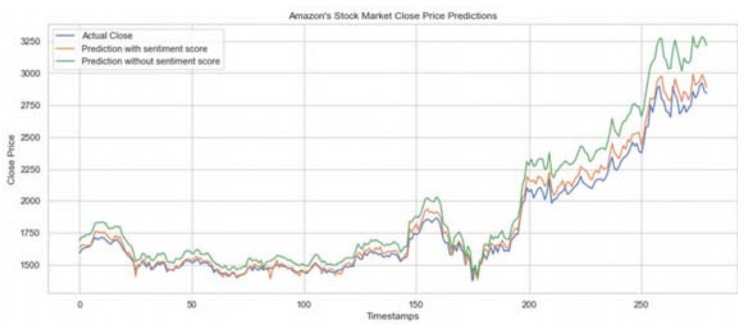

Figure 3. Prediction result comparison for Amazon stock (Close Price)

Authorized licensed use limited to: California State University Fresno. Downloaded on July 01,2021 at 09:46:40 UTC from IEEE Xplore. Restrictions apply.

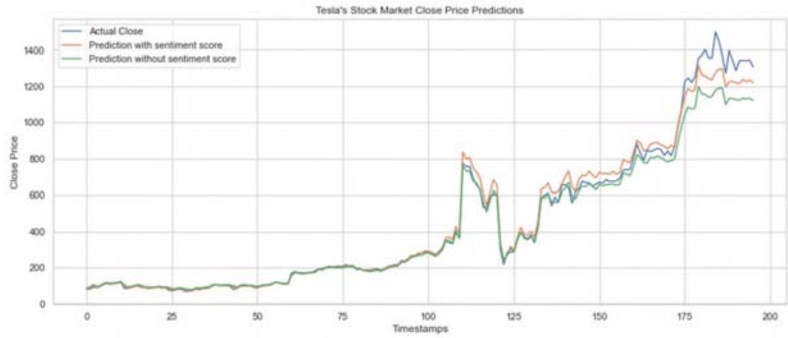

Figure 4. Prediction result comparison for Tesla stock (Close Price)

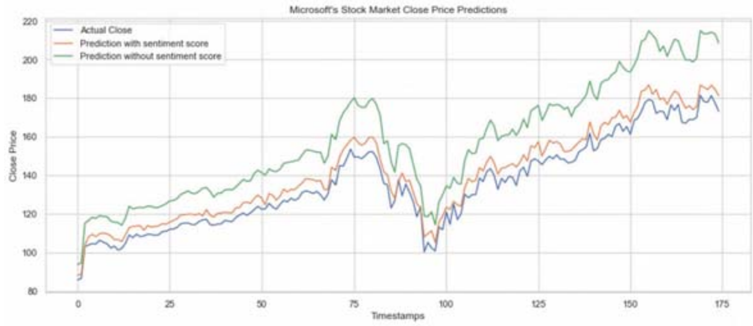

Figure 5. Prediction result comparison for Microsoft stock (Close Price)

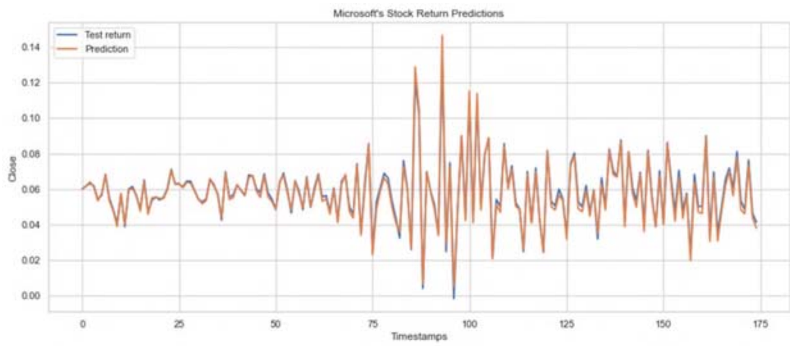

Figure 6. Prediction result for Microsoft stock (Return, with Sentiment Score)

TABLE V. MSE AND MAE VALUES FOR THREE STOCKS (PREDICT DAILY CLOSING PRICE)

MSE for testing data	Amazon	Microsoft	Tesla
With Score	0.00109	0.00497	0.03526
Without Score	0.01112	0.07264	0.08649

MAE for testing data	Amazon	Microsoft	Tesla
With Score	0.02552	0.06627	0.10936
Without Score	0.08518	0.25703	0.13917

TABLE VI. MSE AND MAE VALUES FOR THREE STOCKS (PREDICT DAILY STOCK RETURN)

MSE for testing data	Amazon	Microsoft	Tesla
With Score	0.00379	0.04961	0.15883
Without Score	0.01207	0.05288	0.16293

MAE for testing data	Amazon	Microsoft	Tesla
With Score	0.02461	0.10909	0.18536
Without Score	0.06029	0.12043	0.19679

V. CONCLUSION AND FUTURE WORK

This paper first uses NLP technical tools to analyze and quantify public emotion, and proposes a multilayers LSTM model to predict stock price and stock return, then compares the prediction effect between the model with both news sentiment score and historical stock technical indicators as input and the model based only on historical stock technical indicators. The

Authorized licensed use limited to: California State University Fresno. Downloaded on July 01,2021 at 09:46:40 UTC from IEEE Xplore. Restrictions apply.

experimental results based on three large cap company (Amazon, Microsoft and Tesla) show that compare to model only considers stock time-series data, model used both sentiment score from news article effectively improve the stock prediction accuracy with smaller MSE and MAE value. In addition, from the MSE value for both models, the LSTM predictor shows a good prediction result, which indicates LSTM model is efficient in time-series prediction, such as stock price and stock return. Thus, this paper verifies the significance of public sentiment from news for the stock market forecasting and demonstrates the importance and feasibility of investigating public emotion when conducting stock market prediction and research.

In future work, it is possible to make feature selection for stock technical indicators and select features with high significance into the model, which might improve model quality further. Additionally, sentiments from other news websites could be included in order to make sentiment analysis system more comprehensive. Furthermore, considering more stocks from different industry may be helpful to identify interesting differences between companies for the effect of sentiment analysis on stock prediction.

ACKNOWLEDGEMENT

I would like to appreciate Professor Stephen Coggeshall for his valuable suggestions and persistent guidance.

REFERENCES

[1] E. Fama. The Behavior of Stock-Market Prices. The Journal of Business,1965, 38(1), 34-105.

[2] R. Schumaker, H. Chen. A quantitative stock prediction system based on financial news. Information Processing & Management, 2009, 45(5), 571-583.

[3] G. Mian, S. Sankaraguruswamy. Investor Sentiment and Stock Market Response to Earnings News. The Accounting Review,2012, 87(4), 1357-1384.

[4] B.N. Ashraf. Stock markets' reaction to COVID-19: Cases or fatalities? Research in International Business and Finance, 2020, 54, 101249.

[5] T.H. Nguyen, K. Shirai, J. Velcin. Sentiment analysis on social media for stock movement prediction. Expert Systems with Applications, 2015, 42(24), 9603-9611.

[6] W. Matthias. Reuters sentiment and stock returns. KOF Swiss Economic Institute, ETH Zurich, 2011, 288.

[7] G. Ranco, D. Aleksovski, G. Caldarelli, M. Grčar, I. Mozetič. The Effects of Twitter Sentiment on Stock Price Returns, PLoS ONE, 2015, 10(9), e0138441.

[8] M. Nikou, G. Mansourfar, J. Bagherzadeh. Stock price prediction using DEEP learning algorithm and its comparison with machine learning algorithms. Intelligent Systems in Accounting, Finance and Management, 2019, 26(4), 164-174.

[9] S. Deng, T. Mitsubuchi, K. Shioda, T. Shimada, A. Sakurai.A. Combining Technical Analysis with Sentiment Analysis for Stock Price Prediction. 2011 IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing, Sydney, NSW, 2011, pp. 800-807.

[10] Z. Jin, Y. Yang, Y. Liu. Stock closing price prediction based on sentiment analysis and LSTM. Neural Comput & Applic, 2020, 32, 9713–9729.

[11] G. Bontempi, S.B. Taieb, Y. Borgne. Machine Learning Strategies for Time Series Forecasting. Business Intelligence, 2013, 138.

[12] A. Pak, P. Paroubek. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. Conference: Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23.

[13] C.J. Hutto, E.E. Gilbert.. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14), 2014.

Authorized licensed use limited to: California State University Fresno. Downloaded on July 01,2021 at 09:46:40 UTC from IEEE Xplore. Restrictions apply.