Sentiment Classification of Financial Texts for Stock Forecasting using LSTM Technique

[1]Kanchan Raipure, [2]Prof. MahendraSahare, [3]Prof. Anurag Shrivastava
M. Tech. Scholar[1], Associate Professor[2,3]
Department of Computer Science and Engineering
NRI Institute of Information Science and Technology, Bhopal

Abstract: -The stock market is an emerging network that offers an infrastructure for all financial transactions from the world in a dynamic rate called stock value, which is devised using market stability. Prediction of stock values provides huge profit opportunities which are considered as an inspiration for research in stock market prediction.Long short term memory (LSTM) is a model that increases the memory of recurrent neural networks. Recurrent neural networks hold short term memory in that they allow earlier determining information to be employed in the current neural networks. For immediate tasks, the earlier data is used. We may not possess a list of all of the earlier information for the neural node.The long short-term memory (LSTM) and gated recurrent unit (GRU) models are popular deep-learning architectures for stock market forecasting. Various studies have speculated that incorporating financial news sentiment in forecasting could produce a better performance than using stock features alone. This study carried a normalized comparison on the performances of LSTM and GRU for stock market forecasting under the same conditions and objectively assessed the significance of incorporating the financial news sentiments in stock market forecasting. Both the LSTM-News and GRU-News models are able to produce better forecasting in stock price equally.

Keywords: -Stock Market, LSTM, GRU, Neural Network

I. INTRODUCTION

Sentiment analysis can be performed in many classification algorithms considering a single word (unigram) as a feature. In certain cases unigrams can lead to misclassification of sentiments. New methods were proposed to use the N-gram approach for analysis of sentiments. Certain investigations revealed that as the value of N increases the accuracy of classification decreases. N value was tested with only three as trigrams but it needs more training time. Bigrams were used in some works, which is formed by combining the adjacent words results in a slighter increase in accuracy than unigrams. To eliminate the overhead of input the bigrams are checked with the predefined lexicons to extract the essential bigrams. In the sentiment classification machine learning techniques are widely used at various levels. Presently, the machine learning based approaches are performing on par with human level accuracy (Bayes error) in most of the tasks. Human Generated Classification is a way to evaluate an explanation / review is by asking humans to guess the output of a model based on the explanation and the input. This is referred as forward simulation/prediction [1, 2]. Analysis of sentiments is done at three levels: Document level is used to classify whether the opinion of the document is positive, negative or neutral [3]. The sentence level is used to determine whether the document's opinion is positive, negative or neutral. The aspect level is used to focuses on all expression of sentiments present within a given document. The advantage of learning-based methodologies is that it is very simple and quick to construct. Online reviews are much more efficient and flexible algorithms are required for sentiment classification than the current approaches, which changes rapidly over time. The sentiment sensitive thesaurus has been used to identify the word's sentiment as a lexicon-based approach [4, 5].Indian stock market has shown extreme volatility in the last decade, with prices surging wildly, without any change in fundamentals. The BSE Sensex, which was more than 5000 points in the year 2000, has dipped to less than 3500 points by the second half of the year 2001. In addition to stock market scams at regular intervals, the fall of information technology shares, which has reached to the highest levels, has brought down the Sensex drastically in the financial year 2000-01. Later in 2008 also SENSEX started at 20,352 in January and drastically fell to 9162 in December. The security market exhibits various patterns and styles which are either seasonal at times and shall not follow any seasonality some more times.

II. SENTIMENT ANANLYSIS

The overview of the sentiment analysis process is shown in Figure 1. The objective of sentiment analysis is to extract the opinions from the reviews, identify the sentiment they exhibit and finally to classify the sentiment polarity.

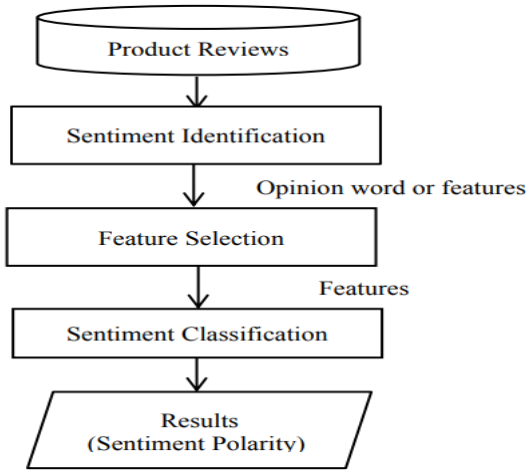

Figure 1: Sentiment classification model

The two main kinds of approaches in sentiment analysis are knowledge-based approaches and machine learning approaches, which are classified from sentiment analysis and classification techniques. Generally, they depend on the analyzing text sentences, identifying the basic sentiment words and then search the dictionary for their relevance [6, 7].

Generative Model for Sentimental Analysis:-In a typical model of sentimental analysis, it defines some models for categorization, which are the generative models. This model examines the joint distribution of all the related data with parameters. These parameters are generally taken as those aspects that reflect the latent structures or properties in the data. It is used to perform sentiment analysis and opinion mining on any type of text data. The generative model for sentimental analysis is categorized into upstream and downstream. Upstream model is used to generate a word as discrete labels and assume that there is different topic proportions labeled under different sentimental labels [8].

In the downstream model, the sentimental variable is assumed to depend on topics, thus the sentiment variable can be regarded as in the downstream and the model attempts to capture how other variables influence the sentiment variable. Upstream model is used to generate the word in the text document. Topic Sentiment Mixture (TSM) model is one of the upstream generative sentimental models. The two additional sentimental models are introduced by the Topic Sentimental Mixture model that is one for positive opinion and another one for a negative opinion [9]. In the machine learning supervised LDA, topics are organized according to their semantic structure and the sentiment labels in the input document are modeled. Most of the nature inspired-algorithms are used only for optimization, so machine learning algorithms are considered for classification improvement in sentiment analysis for the research work [10, 11].

III. STOCK MARKET PREDICTION

Stock market prediction is a significant task for the financial decision-making process and investment. Even though stock price prediction is a key problem in the financial world, it contributes to the growth of efficient methods for stock exchange transactions. Generally, stock markets are in the form of non-stationary, non-linear and uncertain even so financial experts recognized it is complex to produce precise predictions. Stock market prediction is a challenging job due to its high dynamic and unstable. Stock market prediction plans to compute the future value of a company stock trade on exchange as well as consistent prediction of future stock prices obtains high profits to investors. Various researches applied numerical data and news for the prediction of the stock market. Commonly, based on the number of information sources, the stock market prediction technique is experimented on selecting the numerical data by analyzing the news data [12].

In basic, forecasting behaviors are separated into three levels, such as short, medium and long. Furthermore, stock market movements are influenced by various macro economical aspects, like bank exchange rate, commodity price index, investors' expectations, bank rate, general economic conditions, investor's psychology, firms' policies, institutional investors' choices, political events and so on [8, 9]. Additionally, stock value indices are computed using higher market capitalization stocks, whereas several technical parameters are also employed to obtain statistical information about stock price values [13]. In the stock market, there are two assumptions for predicting stock price value. The first one is EMH stating at any time, stock price completely confines all identified information about stock where all identified information's are utilized through market participants and also random price variations obtains new random information's.

Therefore, stock prices execute a random walk, that is every future price does not follow any patterns or trends. This assumption deduces fluctuations, so incomplete or delayed information controls the stock market prices. In addition, an exterior incident influences successive stock market prices, although the precise prediction of a stock price is complex. From the prediction perception, it can be categorized into two types, namely stock price trend and stock price forecast. The stock price trend is also named as classification, and stock price forecast is also termed as regression [14]. Basically, the time duration for stock price trend prediction is highly related with previously selected features [7].

The prediction of stock market future price is very significant for investors, because of the identification of suitable movement of stock price decreases the risk of future trend calculation. The industry, economy and other correlated features are considered to compute the intrinsic value of a company, which helps to forecast stock prices from fundamental analysis method. Stock market decision-making technique is a very complex and significant job because of unstable and complex nature of the stock market. It is necessary to discover a huge quantity of valuable information created through the stock market. In addition, every investor has an imminent

requirement for identifying future behaviors of stock prices.

Although, it helps the investors to achieve the best profit by identifying the best moment to sell or buy stocks. Normally, trading in stock market can be performed electronically or physically. The investor becomes the owner or partnership of a particular company, while an investor obtains a particular company share. Furthermore, financial data of the stock market is very complex in nature, so for predicting stock market behavior is also complex. The stock market prediction helps the investors to take investment decisions by offering strong insights regarding stock market behavior for reducing investment risks.

IV. PROPOSED METHODOLOGY

The market volatility study is more important for policy implications and financial market participants for their future earnings. The Up and Down in the market will add a wedge for the market. The SEBI can improve their reforms of National Stock Exchange to educate the investor in terms of risk involved, return and fluctuation in the market.

In this thesis new solutions that overcome aforementioned challenges in share market prediction strategy adopt the long short term memory (LSTM) technique.

Long Short-Term Memory (LSTM) is one of many types of Recurrent Neural Network RNN, it's also capable of catching data from past stages and use it for future predictions.

In general, an Artificial Neural Network (ANN) consists of three layers:1) input layer, 2) Hidden layers, 3) output layer.

In a NN that only contains one hidden layer the number of nodes in the input layer always depend on the dimension of the data, the nodes of the input layer connect to the hidden layer via links called 'synapses'.

The relation between every two nodes from (input to the hidden layer), has a coefficient called weight, which is the decision maker for signals.

The process of learning is naturally a continues adjustment of weights, after completing the process of learning, the Artificial NN will have optimal weights for each synapses.

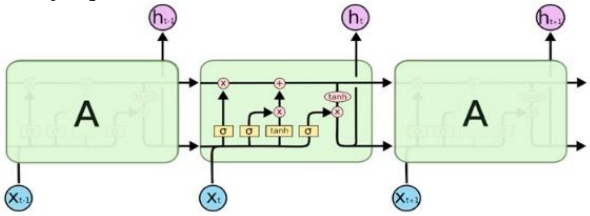

Fig. 1: The internal structure of an LSTM

The principal component of LSTM is the cell state. To add or remove information from the cell state, the gates are used to protect it, using sigmoid function (one means allows the modification, while a value of zero means denies the modification.). We can identify three different gates:

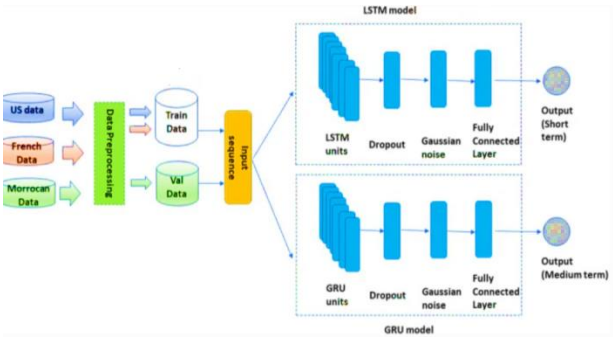

Fig. 2: Flow Chart of Proposed Methodology

Forget gate layer: Looks at the input data, and the data received from the previously hidden layer, then decides which information LSTM is going to delete from the cell state, using a sigmoid function (One means keeps it, 0 means delete it). It is calculated as:

$$f_t = \sigma(w_f[h_{t-1}, x_t] + b_f) \quad (1)$$

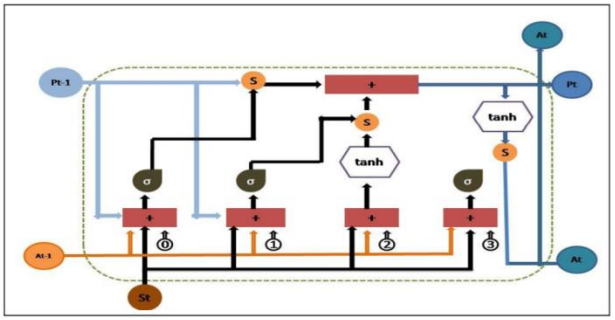

Fig. 3: Working of LSTM

Input/Update gate layer: Decides which information LSTM is going to store in the cell state. At first, input gate layer decides which information will be updated using a sigmoid function, then a Tanh layer proposes a new vector to add to the cell state. Ten the LSTM update the cell state, by forgetting the information that we decided to forget, and updating it with the new vector values. It is calculated as:

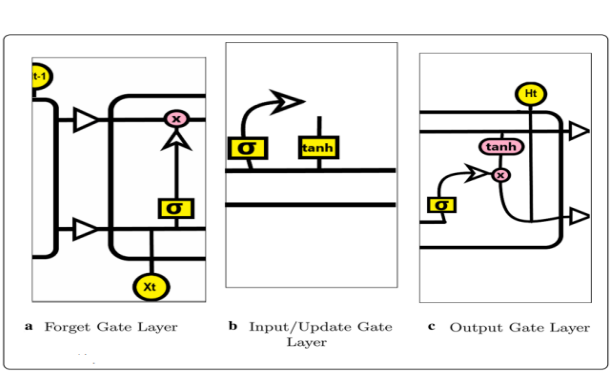

Fig. 4: LSTM Layer

$$i_t = \sigma(w_i[h_{t-1}, x_t] + b_i) \quad (2)$$

and

$$\tilde{C}_t = tanh(w_c[h_{t-1}, x_t] + b_c) \quad (3)$$

Output Layer: decides what will be our output by executing a sigmoid function that decides which part of the cell LSTM is going to output, the result is passed through a Tanh layer (value between − 1 and 1) to output only the information we decide to pass to the next neuron. It is calculated as:

$$O_t = \sigma(w_o[h_{t-1}, x_t] + b_o) \quad (4)$$

and

$$h_t = O_t \times tanh(C_t) \quad (5)$$

Gated Recurrent Unit (GRU):-

Gated Recurrent Unit GRU was introduced in 2014 by Cho et al. To solve the vanishing gradient problem experienced by classical recurrent networks.

Same as LSTM, the input value interacts with the information from the previous state to calculate the different values of intermediate gates which will subsequently be used to decide on the value to be output. GRU is simplified and only update gate (zt) and reset gate (rt) are introduced. In GRU, the update (or input) gate decides how much input (xt) and previous output (ht-1) to be passed to the next cell and the reset gate is used to determine how much of the past information to forget. The current memory content ensures that only the relevant information needs to be passed to the next iteration, which is determined by the weight *W*. The main operations in GRU are governed by the following formulae.

Update gate:

$$z_t = \sigma\left(W_z * [h_{t-1}, x_t]\right)$$

Reset gate:

$$r_t = \sigma\left(W_r * [h_{t-1}, x_t]\right)$$

Table 1: Training parameter data of LSTM and GRU

Model	Sequential – RNN
Type	LSTM, GRU
Hidden Units	7
Input shape	1,1
Verbose	False
Output layer	(TimeDistributed(Dense(1)))
Loss Function	MAE (Mean Absolute Error)
Optimizer	ADAM
Compilation Time	0.01620 S
Total params	260
Trainable params	260
Non-trainable params	0
Epoch	100
Batch size	128

V. SIMULATION RESULTS

Python is a general programming language and is broadly utilized in a wide range of disciplines like general programming, web improvement, programming advancement, information investigation, AI and so forth Python is utilized for this task since it is truly adaptable and simple to utilize and furthermore documentation and local area support is exceptionally huge.

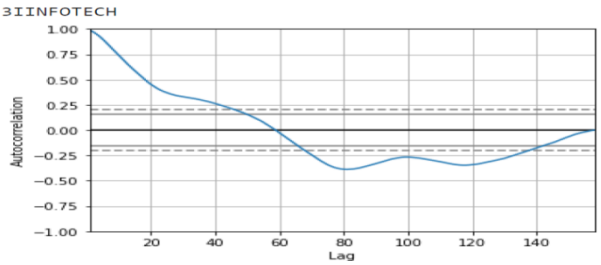

Fig. 5: AC of 3IINFORTECH

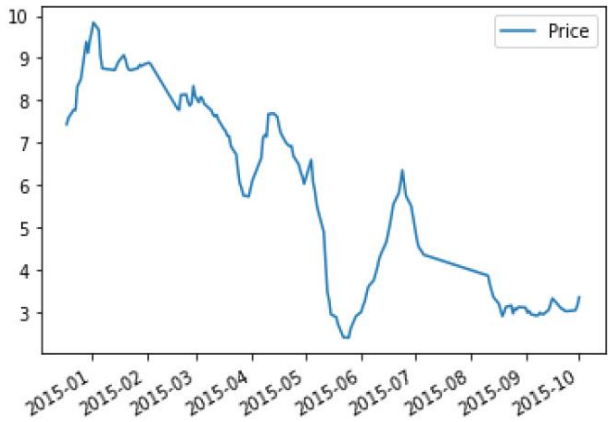

Fig. 6: Price of 3IINFORTECH

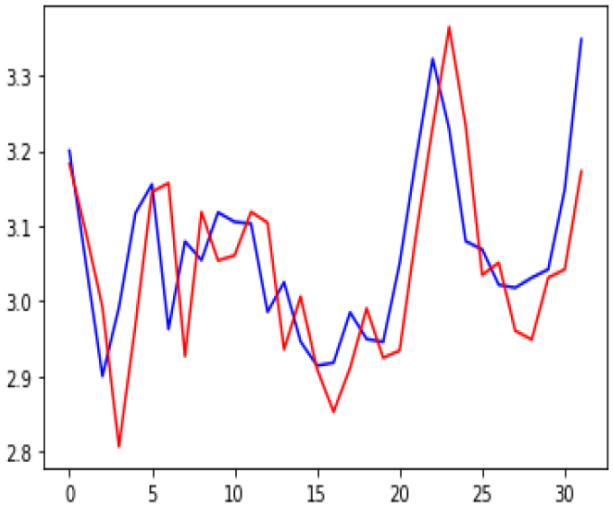

Fig. 7: Prediction and Real value of 3IINFORTECH

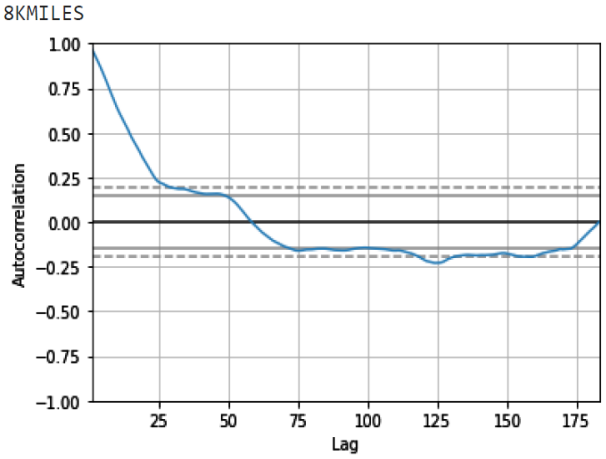

Fig. 8: AC of 8KMILES

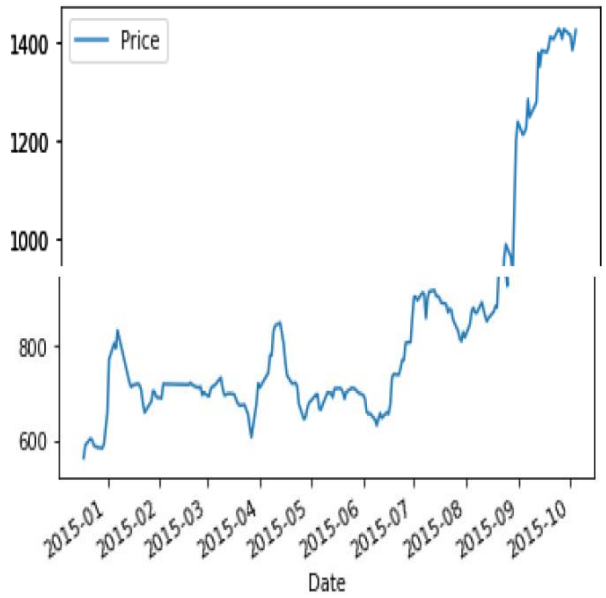

Fig. 9: Price of 8KMILES

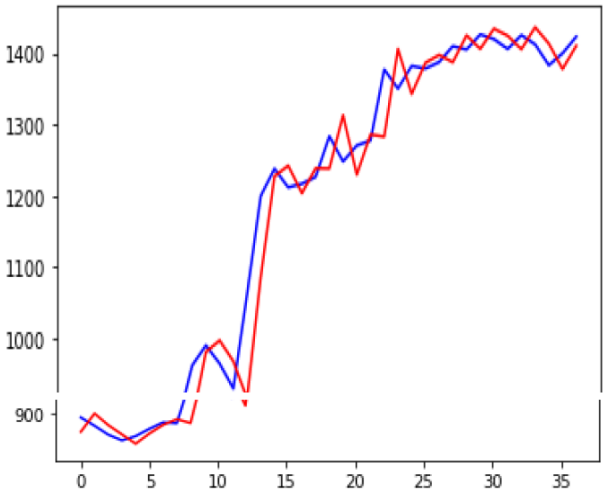

Fig. 10: Prediction and Real value of 8KMILES

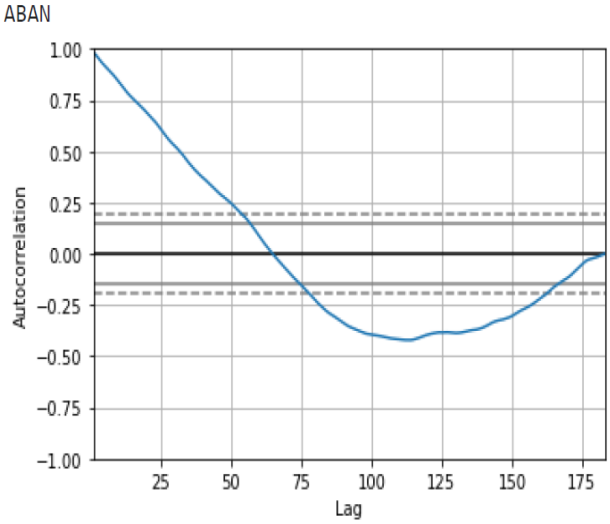

Fig. 11: AC of ABAN

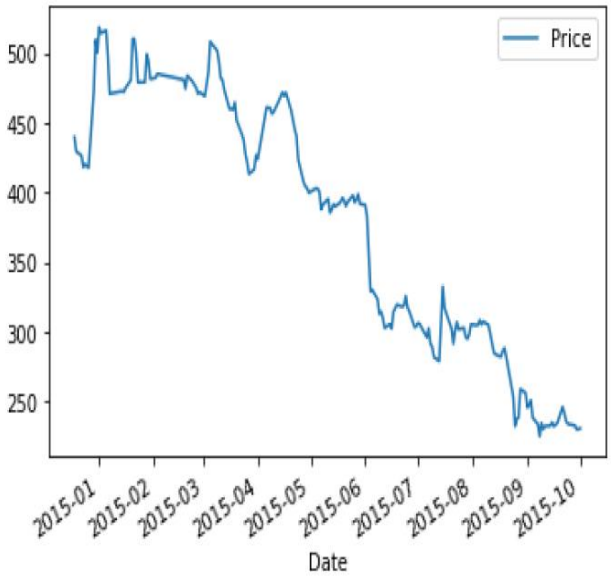

Fig. 12: Price of ABAN

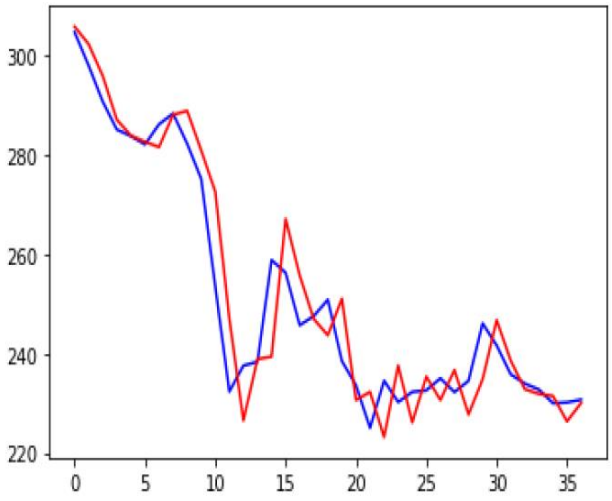

Fig. 13: Prediction and Real value of ABAN

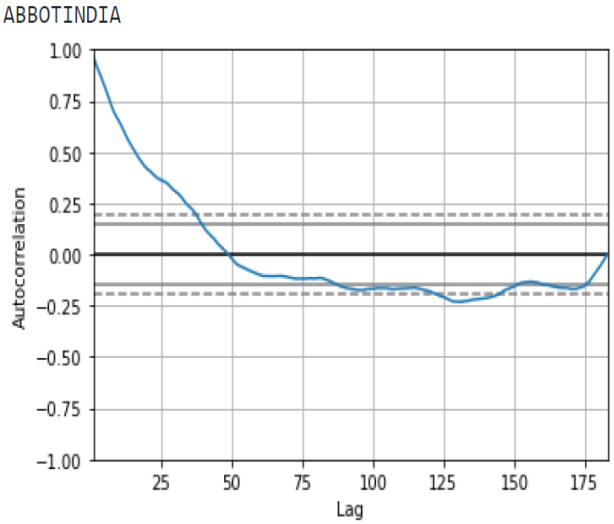

Fig. 14: AC of ABOUTINDIA

Fig. 15: Price of ABOUTINDIA

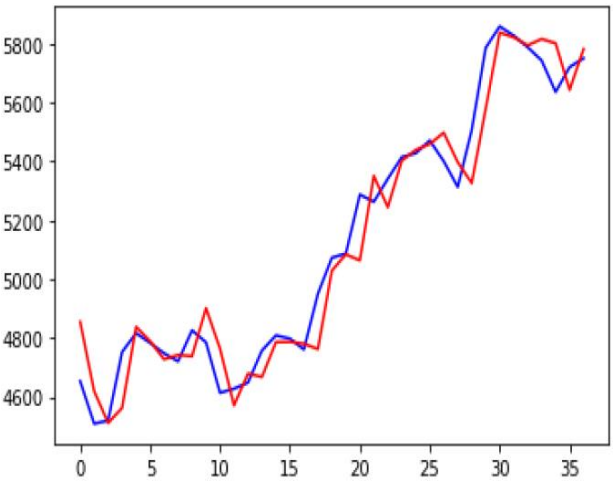

Fig. 16: Prediction and Real value of ABOUTINDIA

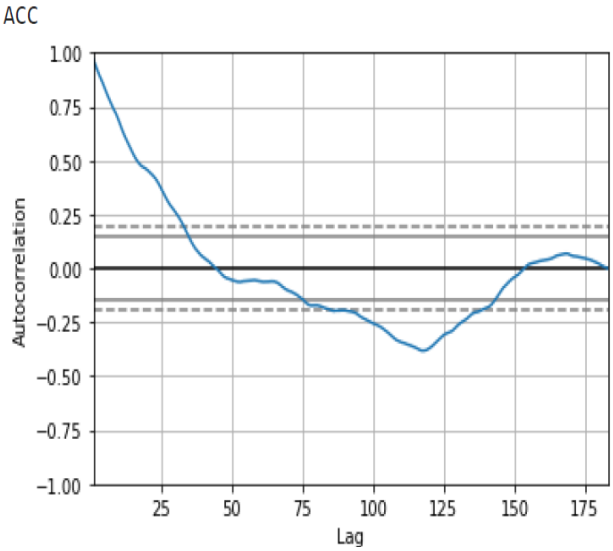

Fig. 17: AC of ACC

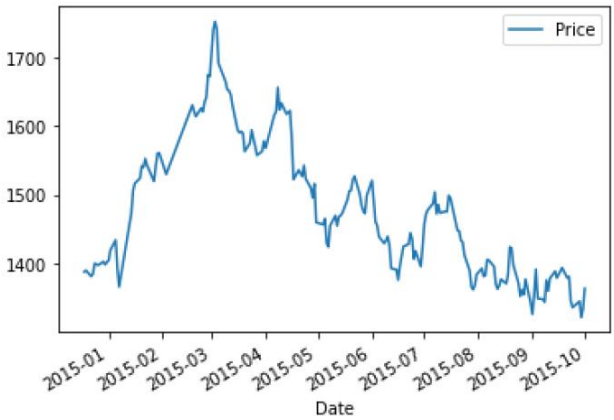

Fig. 18: Price of ACC

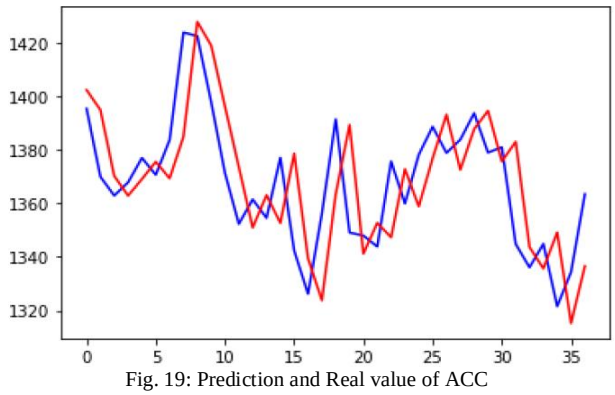

Fig. 19: Prediction and Real value of ACC

VI. CONCLUSION

Finally, a novel approach is proposed to improve the accuracy factor in classification by considering precision and recall which are also important measures in classification. In the final work, the movie review dataset is analysed for sentiment classification with the most frequently used neural network. Here combinations of features are considered and results are analysed for the most commonly used neural network. Unigrams have given better results compared to other n-gram

approaches in this approach. The proposed method combines the previously proposed hybrid approaches with LSTM and GRU for further improvement in classification accuracy and the results are compared with the existing machine learning algorithms. Other performance metrics such as precision and recall are also considered for evaluation with different classifiers.

REFERENCES

[1] Shanshan Dong and Chang Liu, "Sentiment Classification for Financial Texts Based on Deep Learning", Hindawi, Computational Intelligence and Neuroscience, Volume 2021.

[2] Shravan Raviraj, ManoharaPai M M. and Krithika M Pai, "Share price prediction of Indian Stock Markets using time series data - A Deep Learning Approach", IEEE Mysore Sub Section International Conference (MysuruCon), IEEE 2021.

[3] J. J. Duarte S. M. Gonzalez and J. C. Cruz "Predicting stock price falls using news data: Evidence from the brazilian market", Computational Economics vol. 57 no. 1 pp. 311-340 2021.

[4] JingqiLiu;XinzhenPei;Junyan Zou, "Analysis and Research on the Stock Volatility Factors of Chinese Listed Companies Based on the FA-ANN-MLP Model", International Conference on Computer, Blockchain and Financial Development (CBFD), IEEE 2021.

[5] K. M. El Hindi, R. R. Aljulaidan, H. AlSalman, and H. AlSalman, "Lazy fine-tuning algorithms for na¨ıve Bayesian text classification," Applied Soft Computing, vol. 96, p. 106652, 2020.

[6] G. Ding and L. Qin "Study on the prediction of stock price based on the associated network model of lstm" International Journal of Machine Learning and Cybernetics vol. 11 no. 6 pp. 1307-1317 2020.

[7] S. T. Z. De Pauli M. Kleina and W. H. Bonat "Comparing artificial neural network architectures for brazilian stock market prediction" Annals of Data Science vol. 7 no. 4 pp. 613-628 2020.

[8] Y.-T. Tsai, M.-C. Yang, and H.-Y. Chen, "Adversarial attack on sentiment classification," in Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Florence, Italy, 2019.

[9] Z. Zhang, Z. Wang, C. Gan, and P. Zhang, "A double auction scheme of resource allocation with social ties and sentiment classification for Device-to-Device communications," Computer Networks, vol. 155, pp. 62–71, 2019.

[10] Zhihao PENG, "Stocks Analysis and Prediction Using Big Data Analytics", International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), IEEE 2019.

[11] A. Site D. Birant and Z. Isik "Stock market forecasting using machine learning models" 2019 Innovations in Intelligent Systems and Applications Conference (ASYU) pp. 1-6 2019.

[12] A. J. Balaji D. H. Ram and B. B. Nair "Applicability of deep learning models for stock price forecasting an empirical study on bankex data" Procedia computer science vol. 143 pp. 947-953 2018.

[13] A. Dingli and K. S. Fournier "Financial time series forecasting-a machine learning approach" Machine Learning and Applications: An International Journal vol. 4 no. 1/2 pp. 3 2017.

[14] S. Ot´alora, O. Perdomo, F. Gonz´alez, and H.M¨uller, "Training deep convolutional neural networks with active learning for exudate classification in eye fundus images," Lecture Notes in Computer Science, vol. 10552, pp. 146–154, 2017.

[15] Z. Li, "End-to-End adversarial memory network for crossdomain sentiment classification," in Proceedings of the 26th International Joint Conference on Artificial Intelligence, Melbourne, Australia, August 2017.

[16] H. Sagha, N. Cummins, and B. Schuller, "Stacked denoisingautoencoders for sentiment analysis: a review," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 7, no. 5, p. e1212, 2017.