Time Frame Analysis for Sentiment Prediction of Stock Based on Financial News using Natural Language Processing

1st Joy Almeida
Department of Information Technology
Sardar Patel Institute of Technology
Mumbai, India
joy.almeida@spit.ac.in

2nd Kushal Shah
Department of Information Technology
Sardar Patel Institute of Technology
Mumbai, India
shahkushal38@gmail.com

3rd Pratima Singh
Department of Information Technology
Sardar Patel Institute of Technology
Mumbai, India
pratima.singh@spit.ac.in

4th Dr. Rupali Sawant
Department of Information Technology
Sardar Patel Institute of Technology
Mumbai, India
rupali_sawant@spit.ac.in

Abstract - **This research is a study on the impact of a specific stock sentiment based on its news, previous stock movements and finally finding investors sentiment over the stock. This study leverages daily Indian financial news between 2017 and 2021, extracted from various Indian and foreign news sources such as Economic Times, Money Control, Livemint, Business Today, NY Times, WSJ and Washington Post. In this work we propose to analyze news data with a unique pre-processing method that uses vectorization and BERT data processing technology. This is followed by a comparative study and predictive machine learning analysis of following models - Naive Bayes and Recurrent Neural Networks (RNN) with Gated Recurrent Units (GRU), Bi-directional Long Short Term Memory (LSTM) and RNN-LSTM with the pre-processed news data leading us to better accuracy and sentiment findings as compared to other approaches. Based on the comparisons, the results show that - Bi-Directional LSTM layer based on RNN architecture along with BERT Data Processing gives an accuracy of 90.15% leading us to a conclusion of adding a layer of BERT data processing for sentiment analysis to get better results. Further an application feature is being proposed which analyzes real-time stock financial news using RNN-Bi-Directional LSTM, giving a confidence value that is used to calculate overall sentiment of a stock being traded in Indian Stock Exchange for different time frames.**

Keywords— Sentiment Analysis, financial news, Machine Learning, LSTM, Natural Language Processing, RNN, BERT, Data Processing, Stocks

I. INTRODUCTION

Stock market is the place where investors can buy or sell shares. Different investors hold shares of different companies which can be sold anytime in the future. It operates under a specific set of regulations.

Analysis and prediction in stocks have an important role in India's economy. Deciding to buy or sell shares in the stock market is a challenging task for investors. Sentiment of the investors and companies and the change in the sentiments of news have a direct impact on the price. This impact is studied by various authors and researchers as inferred from their work in [11], [12], [13] and [14]. Therefore analyzing the sentiments can help to know the market trend. To be in profit, an investor should buy stocks while the price is low and sell them otherwise. For this the investor must be sure of the current prices which can be estimated using technical analysis. An estimated future price or sentiment will help these investors to get more convincing results.

Different supervised models of machine learning and deep learning have been developed for stock sentiment analysis. However, achieving a higher model accuracy is still difficult. Different factors influence stock market movement, that includes news articles and social media. Our proposed method integrates Recurrent Neural Network (RNN) with Long Short Term Memory (LSTM) to determine the impact of news sentiment on stock market fluctuation and improving the accuracy of volatile stock trend predictions.

In order to help investors identify news-based fundamentally good stocks we have built a web application where investors can get insight on various Indian stock companies. The analysis of a particular stock can be viewed and filtered using the time frame provided by the users. The application offers a comparison between the stock companies as it provides the status of each news (positive/negative) with a confidence value that can range from -1 to 1 and the mean confidence between them for each stock company listed. The proposed application also show analysis of news sentiments for a stock and its impact over different time-frames.

II. LITERATURE REVIEW

In recent times there has been multiple research work being done in the domain of sentiment analysis. The research work being done for sentiment analysis is related to various sectors. In this research we shall keep our focus on sentiment analysis of financial news for a particular stock. The news based sentiment analysis technique using Natural Language Processing is not fully reliable as it has an accuracy of less than 80% and so. Through this work we have tried to find out the reasons for such low accuracy and tried to improve using BERT data processing technique with an LSTM model. The news generated by our proposed model will be helpful in calculating stock sentiment based on time-frame which will help an investor in making their investment decisions in stock markets.

We have exhaustively referred to some of the below mentioned proposed work along with their advantages and disadvantages.

G. Jariwala et. al. [1] researched on how the stock market is driven by various factors, the majority of them being news articles. It focuses on the comparative study of the following models - Support Vector Machine (SVM), K-Mean clustering and Naïve Bayes (NB) in order to determine sentiment of the general market based on news. The dataset used in [1] for training these models were stock news found from moneycontrol.com for the period between 2009 - 2019. The results showed supervised algorithms and more specifically Naive Bayes giving better results over SVM. In this research the data was already cleaned and processed leading them to better results, whereas this doesn't work in actual market conditions, wherein the model will fail in giving accurate results.

In comparison to above Reddit Headlines on Dow Jones Industrial Average of US stock market dataset from 2008-08-08 to 2016-07-01 is used in [2] and [3]. The study by Kameshwari S, Kaniskaa S, Kaushika S and Dr. R. Anuradha [2] centers around the expansion sectors in the share market. This growth is measured by sentiment analysis of news by a comparative study of the following models - Random Forest, Adaboost, Gradient Boosting, Neural Network and XGBoost models. These models were then evaluated based on accuracy score, confusion matrix and F1-score. Results show that only multi model high neural networks give better accuracy nearly 81%, whereas the rest of the models provide accuracy below 70%. To better the accuracy, research work by S. Sridhar and S. Sanagavarapu [3] proposes a neural network based bi-LSTM time-series forecasting model to predict stock prices by using the polarity of the news headlines. The probability is obtained from vectorized headlines fed into ML models acting as events contributing to the price prediction LSTM model. Bi-LSTMs use the LSTM cells to map the dependencies between sequences in both the forward and reverse directions achieving an accuracy of 84.92%.

The drawbacks observed in [2] and [3] majorly were that since there was no restriction added for incoming news and sometimes news may be misunderstood as compared to charts, price-action thus model may fail for real-news. To overcome this, our work uses real-time incoming news from the News API which gives way for exact predictions and provides real-time sentiment analysis of news texts. [5] is a unique text-based unsupervised semantic orientation method using the Noun-Verb approach for sentiment analysis in the financial domain giving better results, accuracy 84%. Using different sentiment indicators, combinations and parts of speech based phrases, seed set phrases are created and sent to a machine learning model that calculates semantic orientation of phrases and expands its knowledge to identify new sets of phrases. This concludes with providing a new set of direction in sentiment analysis by fine tuning and unsupervised classification also giving way for a unique data processing using bidirectional encoding of texts.

In order to understand the time-frame series the work by Yubo Bi, Hanting Liu, Ruiyang Wang and Shiyou Li [4] was studied as it focuses on predicting stock index movement of Dow Jones for different time-frames by developing correlation between index movement and general daily world news headlines. The predictive model algorithms showed accuracy of 28% for SVM, 55% for NB and Random Forest (RF) with a count vectorizer method of the

Natural Language Toolkit achieved an accuracy of 54% - 56%. This indicated a weak correlation between daily world news and index movement of next day, the correlation can be quite evident if the movement is predicted for a longer time-frame as we have studied and demonstrated in our work.

Similarly research work in [8], [9] and [10] is to help prediction of stock prices with better accuracy in different time frames. In these studies researchers have commonly proposed a model to analyze the pattern of stock price movement using LSTM with further modifications. In [8] researchers have proposed a price predictive model based on trained historical stock price data using LSTM. The historical price data is grouped in a bunch of 30, 60 and 90 days for LSTM model analysis. This similar aspect is proposed in our research where news data is grouped in time frames of 30, 60, 90 and 180 days. Paper [9] and [10] are an attempt to find text based sentiment to recognize users' opinions, attitude and emotions about various information on the internet. These analyses provide a deeper assessment of sentiment in industry news bringing a polarity between news and stock prices. The algorithm being used in [10] is a deep RNN and LSTM network processed with words converted into vectors. This Word2Vec aids in tokenizing a unique sequence of words giving deep insights on text meanings and sentiment calculation.

BERT stands for Bidirectional Encoder Representation from Transformers, before performing the experiments we have extensively researched on BERT. Papers [6] and [7] describe the implementation of BERT with LSTM on different datasets and use cases. Research [7] focuses on extracting sarcastic remarks from the news, while research [6] focuses on stock price prediction of Chinese companies using LSTM. BERT is very effective for unlabelled text and has been pre-trained on a vide corpus of text, including entire Wikipedia. BERT being bidirectional learns information from both the left and right sides of a token's context making it more efficient. The working of BERT is described in further sections.

III. PROPOSED METHODOLOGY

The section describes the dataset, preprocessing techniques, and the model used for analyzing the sentiment of the data.

A. Dataset Used

This study uses the data set [15] published by Kaggle which contains the Indian Financial news from Jan 1, 2017, to April 15, 2021. The dataset contains a total of 2,00,500 records with 54% negative and 46% Positive Sentiments. This data is collected from popular Indian news sources like Economic Times, Money Control, Livemint, Business Today, and Financial Express as well as foreign sources like NY Times, WSJ, and Washington Post.

B. Text Pre-processing

First data cleaning is done by deleting the data set's unimportant features. Tokenization is used to convert the headlines to vectors of words. The most commonly used words with very less usage are filtered out using NLTK-stopwords. Then Lemmatization is used to group together the different inflected forms of a word so they can be analyzed as a single item. This data is then split into 30%

testing and 70% training out of which 10% is used for validation.

C. Word Embeddings with BERT

Bidirectional Encoder Representations from Transformer is a transformer-based deep learning model proposed by Google AI researchers. The structure of BERT is shown in Fig. 1. The model pre-trains deep bidirectional representation by joint conditioning on both left and right contexts in all layers. BERT has been trained on a large corpus of text data, which allows it to generate highly expressive and context-aware representations for words and phrases. This makes BERT particularly useful for tasks where understanding the meaning and context of a text is important, such as sentiment analysis.
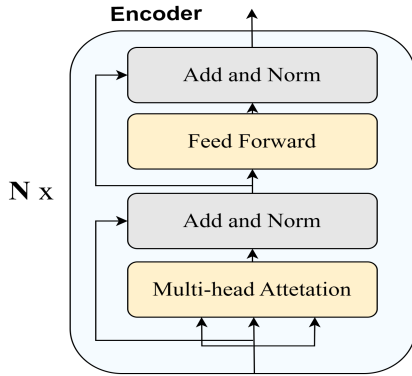
Encoder

Fig 1. Encoder Layer in BERT

We utilized the word embeddings generated from BERT to train the model. By using BERT to generate the word embeddings, our model was able to take advantage of the rich, context-aware representations learned by BERT, which has been shown to be quite effective in a wide range of NLP(Natural Language Processing) tasks.

D. Sentiment Analysis Models

1) Naive Bayes

Naive Bayes theorem is dedicated to extract subjective emotions and feelings from news text. To predict whether a news is positive or negative we have used the famous Bayes Theorem as shown in the below equation.

$$P(J|K) = \frac{P(K|J) \cdot P(J)}{P(K)}$$

P (J|K) stands for probability of news J being positive given K is also positive.

Since, we need to consider more than 2 data points for accurate sentiment measurement, the bayes theorem is modified by assuming each data point as independent. The new equation for probability calculation becomes -

$$P(y \mid x_1 \ldots\ldots x_n) = \frac{P(y) \prod\limits_{i=1}^{n} P(x_i|y)}{P(x_1 \ldots\ldots x_n)} \quad \text{where y, } x_1 \ldots$$

x_n are news data vectors, whose probability of positivity is needed to be calculated.

2) Recurrent Neural Network (RNN)

RNN is capable of processing large sequences of data such as a news document consisting of a large sequence of words. These inputs are indexed at a time step with an ordered list of vectors $v_1, v_2, v_3 \ldots\ldots v_T$. These vectors are aligned or unaligned as per the corresponding target eg - parts of speech step - by - step correspondence.

As a stock trader making strategic investments predicting short-term and long-term sentiments using history of available news, the trader is interested in calculating probability distribution, expected value and variance.

RNN sequencing model reduces language modeling to an autoregressive predictive model. This model decomposes joint density of a sequence resulting in a satisfying first order markov model. This model counts the number of times each word has occurred in each context. The RNN based joint probability equation becomes -

$$P(v_1, \ldots\ldots v_T) = P(v_1) \prod_{t=2}^{T} P(v_t|v_{t-1})$$

3) RNN with Grated Recurrent Units (GRU)

The working of RNN [15] and [18] is that it uses previous state information in series to produce current output.Thus previous output is stored in many timesteps whereas it becomes difficult for RNN to preserve longer timesteps. This flaw overcomes the network. The equation of GRU and LSTM wherein both the cases have memory cells usage to store the activation value of preceding words in the long series.

The GRU architecture is same as RNN with difference in operations maintained by primarily two gates -

a. Reset Gate responsible for Short term memory with equation as -

$$r_t = \sigma (x_t * U_r + H_{t-1} * W_r$$

x_t = input

H_{t-1} = hidden state from previous time step (t-1)

$U_r \text{ and } W_r$ = weighted matrices for reset gate

r_t = ranges from 0 to 1 because of sigmoid function

b. Update Gate for Long Term memory with the equation same as above, having U_u and W_u as weighted matrices for update gate.

The most important part in GRU working is to find candidate hidden states using reset gate and consequently the current hidden state using update gate. If value of r_t is equal to 1 then it means entire information of previous hidden state H_{t-1} is being considered and value of r_t is 0 then previous state is ignored. Using these GRU network gates with RNN, accuracy if the model is calculated for finding sentiment in financial news.

4) Recurrent Neural Network Model with Long Short Term Memory (RNN-LSTM)

A RNN that is designed to handle sequential data, such as time series, text, and speech. LSTMs are a variation of traditional RNNs that are designed to overcome the problem of vanishing gradients, which

occurs when training RNNs on Long term dependencies as inferred from [17].

An LSTM network consists of a series of memory cells, each of which has a set of gates that control the flow of information in and out of the cell. These gates include

1. Input gate: Acts as a gatekeeper for the flow of fresh information into the cell.
2. Forget gate: Manages the flow of information out of the cell.
3. Output gate: Responsible for controlling information out of the cell to the rest of the network.

The structure of LSTM is shown in Fig. 2. The LSTM network uses these gates to selectively remember or forget information from the input sequence, allowing it to learn long-term dependencies in the data.

In our proposed study, the fixed-sized numerical representations generated for each token of input text by BERT are fed to the LSTM model. The model trains on the labeled data to predict the sentiment of new input text.
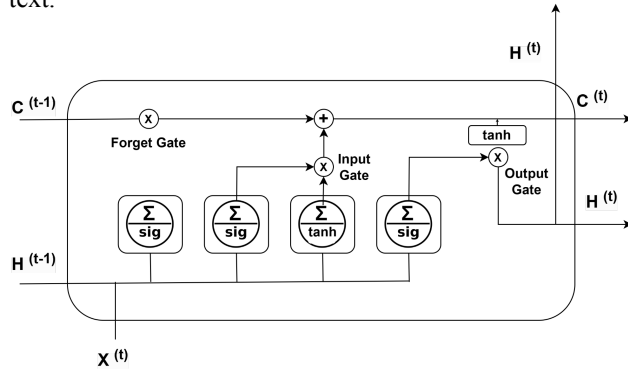

Fig 2. Structure of LSTM Neuron

IV. APPLICATION DESIGN
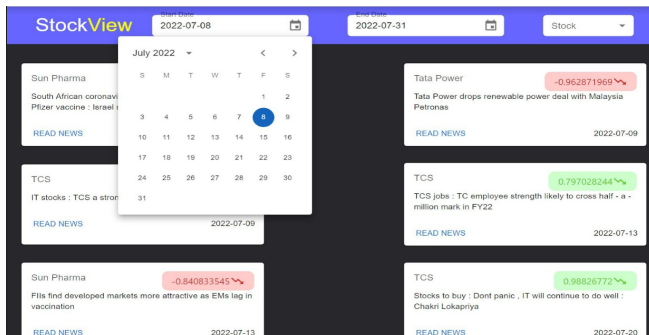
A. User Interface

1) Home Page

Fig 3. Screenshot of the Home Page of the application

Currently the application developed supports only 5 stocks. The home page contains the initial comparison of all 5 stocks based on their positive and negative score in a graphical representation in the form of a bar chart. The home page also contains the news from all 5 stocks with their sentiment value attached to it. In addition to this, the application provides an interface for the user to analyze the data based on the selected time frame.

A user is also provided an option to select a particular stock through which he/she can view additional details of the stock selected.
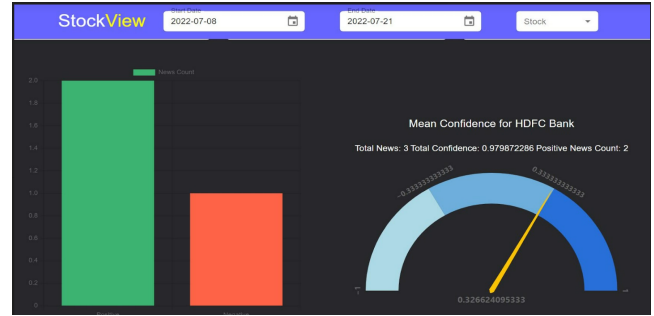
2) Stock Page

Fig 4. Stock Specific Page of the application

The stock page contains the details for a particular stock in more detail. On the top, it displays all information and current news related to the stock. Following this, it shows the current sentiment of the stock in a graphical representation with a needle and a bar graph. This can help the user with more in-depth information about the stock.

B. Data Flow

The news of each stock is fetched using a third-party API and given to the model for pre-processing and sentiment prediction. The data is then saved in the database and fetched using REST API. The structure of the proposed system is shown in Fig. 5.
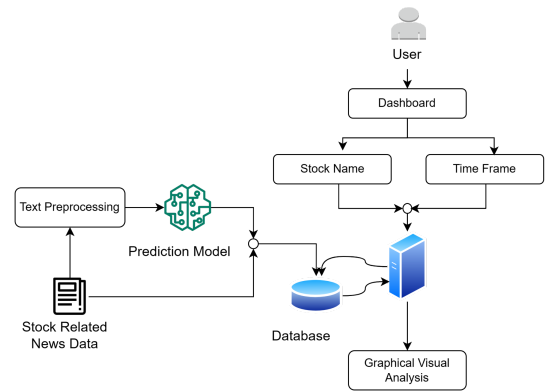

Fig 5. Flow Diagram of the Proposed System
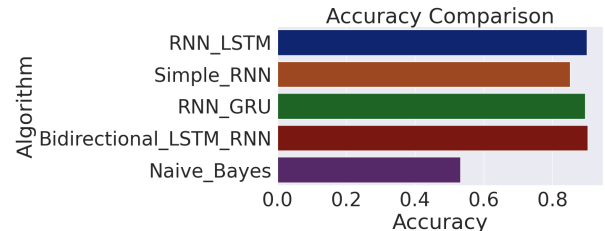
V. RESULTS AND DISCUSSION

Fig 6. Accuracy Comparison with Bar graph

RNN LSTM (Long Short Term Memory) Model		
	Precision	Recall	F1 Score
Class 0	0.91	0.90	0.91
Class 1	0.89	0.89	0.89

Simple RNN (Recurrent Neural Network)		
	Precision	Recall	F1 Score
Class 0	0.87	0.85	0.86
Class 1	0.83	0.85	0.84

RNN GRU (Gated Recurrent Units) Model		
	Precision	Recall	F1 Score
Class 0	0.88	0.93	0.90
Class 1	0.91	0.85	0.88

Bidirectional RNN LSTM Model		
	Precision	Recall	F1 Score
Class 0	0.91	0.91	0.91
Class 1	0.90	0.89	0.89

Naive Bayes		
	Precision	Recall	F1 Score
Class 0	0.55	0.79	0.64
Class 1	0.48	0.23	0.31

Table 1: Classification Reports of Models Implemented

Model	Accuracy
RNN	0.850
RNN LSTM	0.899
RNN GRU	0.894
Bidirectional RNN LSTM	0.901
Naive Bayes	0.532

Table 2. Accuracy Comparison of Models Implemented

The results of these experiments show that Recurrent Neural Networks (RNNs) perform well on the task at hand, with an accuracy of 85.0%. The use of Gated Recurrent Unit (GRU) variants of RNNs and Long Short-Term Memory (LSTM) further improves performance, with accuracy scores of 89.9% and 89.4%, respectively. The highest accuracy was achieved using a Bidirectional RNN with LSTM, which achieved an accuracy of 90.1%. A naive Bayes classifier was also tested, with an accuracy of 53.2%, which is significantly lower than the RNN models. These results suggest that RNNs and their variants are well-suited for this task, and that the bidirectional LSTM RNN is the best model among the ones tested.

VI. Conclusion And Further Work

In this study, we proposed an LSTM-RNN-based system for visually analyzing stock trends based on sentiment analysis on stock-specific news belonging to a particular time frame. Our proposed method integrates Recurrent Neural Network (RNN) based LSTM(Long Short Term Memory) to identify the influence of financial news sentiment on a particular company's stock being traded in stock exchange and forecast volatile stock investment trends more accurately.

In terms of data, both LSTM and Bidirectional LSTM can handle sequential data, such as text, speech, and time-series data. However, Bidirectional LSTM is better suited here because the order of the input sequence is important and bidirectional context is useful.. On the other hand, LSTM may be more appropriate for tasks where the input sequence has a temporal aspect, such as time-series analysis, where the sequence is one-dimensional and only one direction of processing is needed.

It's also worth noting that Bidirectional LSTM may be computationally more expensive than LSTM due to the bidirectional processing. However, with advancements in hardware and software, this may not be a significant issue in many applications.

We also used different models for our comparisons like RNN, RNN GRU, Bidirectional RNN LSTM, and Naive Bayes. As shown in fig 7.1 and fig 7.2, the highest accuracy was achieved by RNN LSTM.

As RNN LSTM networks do better jobs in handling complex sequential data and capturing the long-term dependencies that are present in many natural languages and time series tasks; they provide better accuracy as compared to any other models.

In addition to this, we also developed an application that provides an interface for the user to analyze the stock data in a graphical format based on the selected time frame. This web application is not limited to any platform and can cater needs of an individual irrespective of the area he/she is in.

As a part of future scope we can scale the number of stocks being analyzed by the model. This can in turn give a wider access to sentiment in stocks listed in Stock Exchange of India. It in turn can help lakhs of passive investors investing in the stock market and do not have time to track news of each and every stock.

References

[1] Jariwala, Gaurav Agarwal, Harshit Jadhav, Vraj. (2020). Sentimental Analysis of News Headlines for Stock Market. 1-5. 10.1109/INOCON50539.2020.9298333.

[2] K. S, K. S, K. S and R. Anuradha, "Stock Trend Prediction Using News Headlines," 2021IEEE India Council International Subsections Conference (INDISCON), 2021, pp. 1-5, doi: 10.1109/INDISCON53343.2021.9582219.

[3] S. Sridhar and S. Sanagavarapu, "Analysis of the Effect of News Sentiment on Stock Market Prices through Event Embedding," 2021 16th Conference on Computer Science and Intelligence Systems (FedCSIS), 2021, pp. 147-150, doi: 10.15439/2021F79.

[4] Y. Bi, H. Liu, R. Wang and S. Li, "Predicting Stock Market Movements Through Daily News Headlines Sentiment Analysis: US Stock Market," 2021 2nd International Conference on Big Data Artificial Intelligence Software Engineering (ICBASE), 2021, pp. 642-648, doi: 10.1109/ICBASE53849.2021.00127.

[5] Anita Yadav, C K Jha, Aditi Sharan, Vikrant Vaish,Sentiment analysis of financial news using unsupervised approach,Procedia Computer Science, Volume 167, 2020, Pages 589-598, ISSN 877-0509, https://doi.org/10.1016/j.procs.2020.03.325.

[6] X. Weng, X. Lin and S. Zhao, "Stock Price Prediction Based On Lstm And Bert," 2022 International Conference on Machine Learning and Cybernetics (ICMLC), Japan, 2022, pp. 12-17, doi: 10.1109/ICMLC56445.2022.9941293.

[7] H. Liu and L. Xie, "Research on Sarcasm Detection of News Headlines Based on Bert-LSTM," 2021 IEEE International Conference on Emergency Science and Information Technology (ICESIT), Chongqing, China, 2021, pp. 89-92, doi: 10.1109/ICESIT53460.2021.9696851.

[8] S. Srivastava, R. Tiwari, R. Bhardwaj and D. Gupta, "Stock Price Prediction Using LSTM and News Sentiment Analysis," 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2022, pp. 1660-1663, doi: 10.1109/ICOEI53556.2022.9776789.

[9] R. Kumar, C. M. Sharma, V. M. Chariar, S. Hooda and R. Beri, "Emotion Analysis of News and Social Media Text for Stock Price Prediction using SVM-LSTM-GRU Composite Model," 2022 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), 2022, pp. 329-333, doi: 10.1109/CISES54857.2022.9844375.

[10] J. S. Vimali and S. Murugan, "A Text Based Sentiment Analysis Model using Bi-directional LSTM Networks," 2021 6th International Conference on Communication and Electronics Systems (ICCES), 2021, pp. 1652-1658, doi: 10.1109/ICCES51350.2021.9489129.

[11] S. Mohan, S. Mullapudi, S. Sammeta, P. Vijayvergia and D. C. Anastasiu, "Stock Price Prediction Using News Sentiment Analysis," 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService), 2019, pp. 205-208, doi: 10.1109/BigDataService.2019.00035.

[12] P. Ray, B. Ganguli and A. Chakrabarti, "A Hybrid Approach of Bayesian Structural Time Series With LSTM to Identify the Influence of News Sentiment on Short-Term Forecasting of Stock Price," in IEEE Transactions on Computational Social Systems, vol. 8, no. 5, pp. 1153-1162, Oct. 2021, doi: 10.1109/TCSS.2021.3073964.

[13] M. Omarkhan, G. Kissymova and I. Akhmetov, "Handling data imbalance using CNN and LSTM in financial news sentiment analysis," 2021 16th International Conference on Electronics Computer and Computation (ICECCO), 2021, pp. 1-8, doi: 10.1109/ICECCO53203.2021.9663802.

[14] Y. Guo, "Stock Price Prediction Based on LSTM Neural Network: the Effectiveness of News Sentiment Analysis," 2020 2nd International Conference on Economic Management and Model Engineering (ICEMME), 2020, pp. 1018-1024, doi: 10.1109/ICEMME51517.2020.00206.36

[15] Dataset source on Indian Financial News from Kaggle: https://www.kaggle.com/datasets/harshrkh/india-financial-news-headlines-sentiments