International Journal of Research in Engineering and Applied Sciences(IJREAS)
Available online at http://euroasiapub.org/journals.php
Vol. 12 Issue 12, December- 2022
ISSN(O): 2249-3905, ISSN(P) : 2349-6525 | Impact Factor: 8.202 |
Thomson Reuters ID: L-5236-2015

Sentiment analysis and deep learning are used to predict stock prices for Indian markets

Jatinder Kaur
Associate Professor, Department of Mathematics, Guru Nanak Girls College (of Affiliation)
Yamunanagar, Haryana, India
Email id: jatinderkaur.gng@gmail.com

Abstract: Predicting the stock market has been a busy field of study for a considerable amount of time. The initiation of totalling and machine knowledge has augmented research and created new opportunities. Our goal in this search study was to forecast share price movement in the future by utilising historical pricing and sentiment data that was readily available. The exercise involved the employment of two models, the first of which was the LSTM model, which used historical prices as the independent variable. Major parameters for the Random Forestry Model employed in the second portion included gold prices, oil prices, the exchange rate between the US dollar and the Indian rupee, and sentiment analysis obtained with the Intensity Analyser. Additionally, securities yields were incorporated into the algorithm to increase its accuracy. As the final result, the prices of few stocks—Reliance, TCS, and SBIs were forecasted through the use of the two models mentioned above. The RMSE metric was used to assess the outcomes.

Keywords: Sentiment analysis, Random Forest

- **Introduction**

Predicting future stock values by machine learning and other forms of artificial intelligence has stood the aim of this exercise. The exercise began with a thorough analysis of the body of work that has been published in this ground. A short list of the study articles and online resources that address this issue has been supplied as references. The EMH and the Random Walk were the foundations of early research on stock market prediction. Several studies, including those by Gallagher, Kavussanos, and Butler, demonstrate that stock market prices can be anticipated to a degree. A further hypothesis that is being investigated is if it is

possible to forecast changes in economic and commercial indicators using early indicators that are taken from internet sources (blogs, twitter feeds, etc.). Similar analysis has been conducted in other research domains; for example, Gruhl (2005) the relationship amid online conversation movement. Mishne (2006) have employed blog mawkishness analysis to forecast box office receipts. The connection between flouting financial news and variations in stock price was examined by Schumaker (2009). Bollen (2011) conducted one of the most significant studies in the subject of stock prediction, looking into the association among the Dow Jones Industrial Index and public sentiment. Using Twitter (happy, calm, and anxious) were determined. By looking at and categorising the Twitter feeds, Chen (2011) were able to extract investment strategies. After analysing the Twitter stream, Bing et al. came to the conclusion that stock prediction was industry-specific. Negative social network emotions and the DJIA index showed a strong negative association, according to Zhang (2013). In their research, Pagolu (2016) found a significant relationship between the fluctuations in a company's stock price and the sentiments expressed by the general public on Twitter. Their work focused on creating a sentiment analyser to classify tweets into three categories: positive, negative, and neutral, as opposed to utilising a typical word embedding model. In their study, Mittal et al. attempted to use Twitter sentiment analysis to create a tool for managing portfolios. Using DJIA, they examined and evaluated their model. A day ahead of time, the greedy strategy-based model was able to forecast the Buy/Sell choices for the DJIA holdings based on sentiment research of social media. on their study, Chen et al. compared the LSTM model with the Random estimation model, validating the LSTM model's greater accuracy. The LSTM model was built on the basis of the algorithm, and it was used to forecast the direction of pillories on the Chinese Typical Exchange. A study by Tekin (2018) used a variety of forecasting algorithms to evaluate data from 25 of the top corporations. Their research demonstrated the greater significance of the Random Forest approach. Malandri (2013) use the random forest classifier and LSTM multi-layer perceptron (MLP) in their portfolio allocation model. According to an analysis of NYSE data, LSTM produces superior experimental outcomes. Their research combined a number of word embedding and deep learning models to find the combination that might forecast stocks with the highest accuracy. To categorise the information as good or bad, they employ data labelling. Using

International Journal of Research in Engineering and Applied Sciences(IJREAS)
Available online at http://euroasiapub.org/journals.php
Vol. 12 Issue 12, December- 2022
ISSN(O): 2249-3905, ISSN(P) : 2349-6525 | Impact Factor: 8.202 |
Thomson Reuters ID: L-5236-2015

Twitter data as the foundation, the Word2Vec inserting model in blend with LSTM produced the highest typical accuracy athwart the 9 stocks under consideration. The public domain contained stock-related gen at the beginning of the exercise. Information about stocks was sourced from Ya-hoo Finance. The collected data included the standard data points Open, Low, High etc. that are used in stock analysis. For the EDA exercise, data from January of 2007 was utilised. We discovered Domain Exploration through the model-building process, which went beyond the topics already discussed in the Literature Review section. Research was conducted on a variety of macro/global economic and other fundamental criteria in a variety of fields, including finance, budget, trade, and supplementary core manufacturing aspects. The goal was to come up with a final set of criteria that would really affect stock prices. In the end, the following macro parameters were selected as part of the exercise: (which are expected to have a negative relationship with market returns); which is a proxy for fuel and has a significant impact on almost all monetary indicators; government bond yields, which increase economic pressure and impact market returns; and the USD-INR exchange rate, which has a noteworthy impact on innumerable command parameters and is probable to aid in the improved explanation and calculation of movements in stock prices. Sentiment analysis was used in the exercise, which was another important component. It was anticipated that Tweets would serve as the feed aimed at our Mawkishness Analyser; but, due to changes in Twitter policies, obtaining feed data from the social media platform proved to be an insurmountable challenge.

Alternatively, news headlines were sourced manually from a variety of publicly accessible data news fonts, plus BSE, News18, Mint, and so on. As part of the effort, data covering the period from June 1, 2019, to June 28, 2021, was compiled. The information was collected into an excel file with distinct rows representing each news headline and was made accessible on a diurnal basis from the websites mentioned above. Preparing the data was necessary because it was sourced from news websites. To get the data in a format that the Sentiment Intensity Analyser could use, stop word removal exercises, special character removals, and other common pre-processing tasks were completed. As part of the cloud, the intensity analyser displays four different sorts of sentiments: compound, negative, neutral, and positive. While many news items corresponded to a single day, price data has been taken into consideration and is projected daily.

International Journal of Research in Engineering and Applied Sciences(IJREAS)
Available online at http://euroasiapub.org/journals.php
Vol. 12 Issue 12, December- 2022
ISSN(O): 2249-3905, ISSN(P) : 2349-6525 | Impact Factor: 8.202 |
Thomson Reuters ID: L-5236-2015

In order to provide the Mawkishness Analyser with the appropriate sentiment fog aimed at the given date, all news items that corresponded to a given day stood concatenated as a single text input utilising the Random Forest Model, predictions are made utilizing this sentiment cloud in conjunction with the historical Adjacent Price and supplementary macro characteristics previously mentioned.

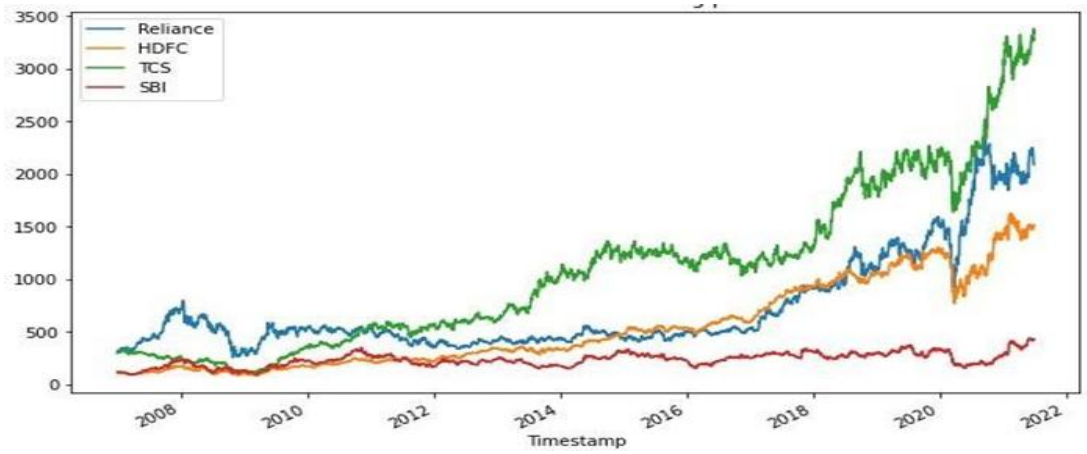

Fig. 1: Historic Performance of Pillories

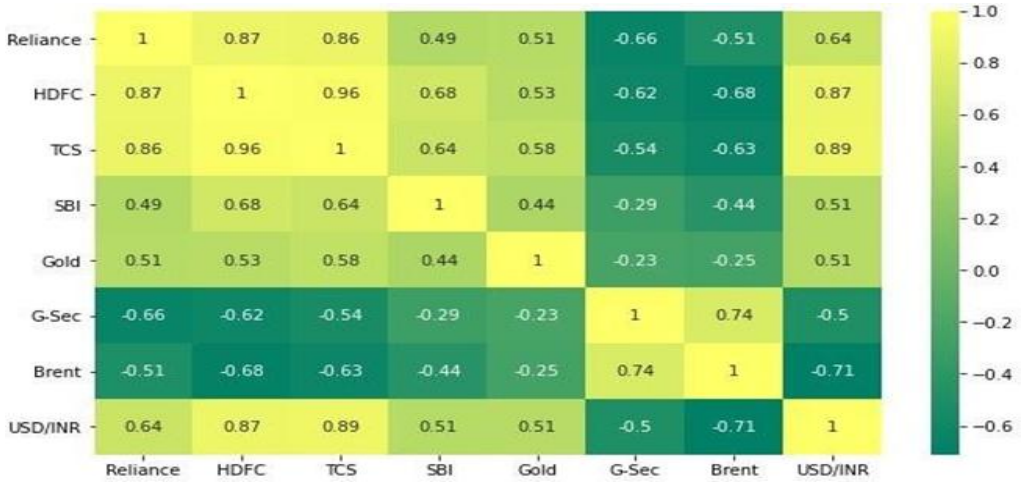

Fig. 2: Correlation - Macro Strictures

International Journal of Research in Engineering and Applied Sciences(IJREAS)
Available online at http://euroasiapub.org/journals.php
Vol. 12 Issue 12, December- 2022
ISSN(O): 2249-3905, ISSN(P) : 2349-6525 | Impact Factor: 8.202 |
Thomson Reuters ID: L-5236-2015

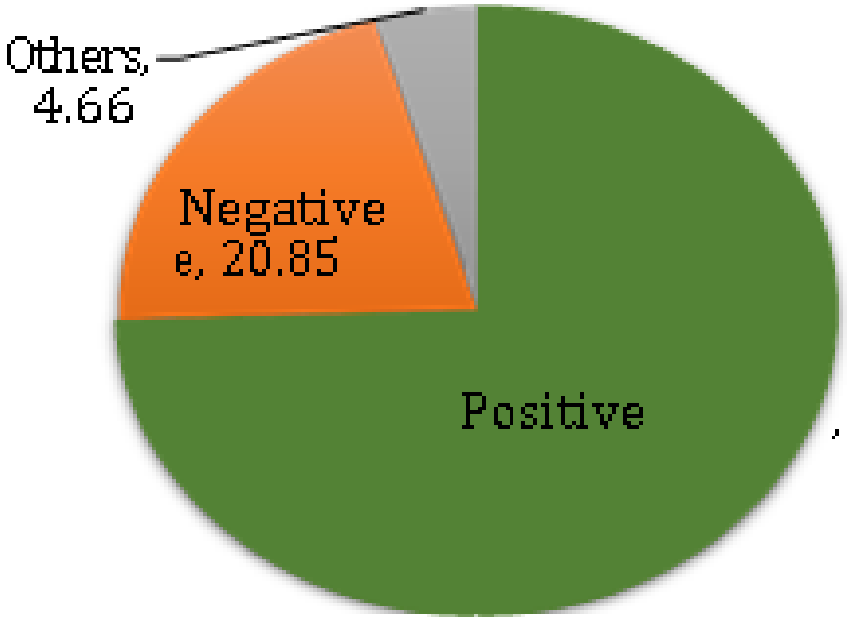

Fig. 3: Instantaneous of Sentiments aimed at Reliance

The table below shows a summary of the several data sources, the range of data, and the data sources

Table 1: Data – Features, Assortment and Sources

Independent Features	**Indicator Used**	**Date Range**	**Data Source**
Gold Prices	Comex 100oz Gold Price	2006 -21	Public Domain
Fuel Prices	Brent Crude Oil Benchmark	2006-21	Public Domain
Bond Yields	10 year Govt. of In-dia Bond Yield	2006-21	Public Domain
Exchange Rate	USD-INR Exchange Rate	2006-21	Public Domain
Sentiment Data	News from websites related to stocks	2019-21	Public Domain
Dependent Variable Close Price	Close price	22006-21	Yahoo Finance

International Journal of Research in Engineering and Applied Sciences(IJREAS)
Available online at http://euroasiapub.org/journals.php
Vol. 12 Issue 12, December- 2022
ISSN(O): 2249-3905, ISSN(P) : 2349-6525 | Impact Factor: 8.202 |
Thomson Reuters ID: L-5236-2015

- **Step by step walk complete of the explanation**

Following the data pre-processing and EDA exercises, an attempt was made to apply a number of machine learning methods in order to reach the necessary error levels. Several algorithms were tested, including Prophet, Arima, KNN, Random Forest Regressor, and Linear Regression. Below is a brief explanation of the algorithms and the parameters taken into account:

1. The system parameters for linear regression are determined by the price of a stock and the time period, which makes the technique widely applicable.
2. The K-nearest neighbour algorithm is a member of the family of algorithms that enable the identification of patterns. It is a subset of lazy learning, often known as instance-based learning.
3. Because classification relies on distance, normalisation is essential to increasing accuracy.
4. Our prediction model made use of the Euclidean distance metric. After testing a neighbour value range of 2 to 9, Greed Search was utilised to get the ideal n value. Five cross-validations were also performed for the hyper-parameters.
5. Auto regressive models resemble regression models, however in this instance, the dependent variable with a particular lag serves as the predictor. Before analysing non-stationary data, the ARIMA Model transforms it into stationary data. 50 functions with order 0 AR, order 1 differencing, and order 1 MA are to be evaluated using the "lbfgs" approach. The model's seasonal component for the AR is 2, its differences are 1, its MA is 0, and its periodicity is 12.
6. Prophet is a method for predicting time series data. It is built on an additive model that fits non-linear trends with seasonality that occurs on a monthly, weekly, and daily basis in addition to holiday impacts.
7. It functions best with multiple sea- sons of historical data and time series with significant seasonal effects.
8. To forecast the daily stock price, we employed a linear curve. Two columns were needed in the data frame that Prophet used: (i) "ds," which was used to record date time series, and (ii) "y," which was used to store the matching values of the time series (stock values).

International Journal of Research in Engineering and Applied Sciences(IJREAS)
Available online at http://euroasiapub.org/journals.php
Vol. 12 Issue 12, December- 2022
ISSN(O): 2249-3905, ISSN(P) : 2349-6525 | Impact Factor: 8.202 |
Thomson Reuters ID: L-5236-2015

9. By utilising historical close price data, LSTM, including Bidirectional-LSTM, was attempted to forecast stock prices. Additionally, hyperparameter optimisation was done to get the best possible outcomes.

The table below shows that LSTM is outperforming the other models that were tried during the study. As a result, LSTM was chosen to forecast stock prices for HDFC Bank, SBI, and TCS, among other corporations. The exercise's next step was sentiment analysis utilising news headlines.

An effort was made to analyse the news data's sentiments. Using the Intensity Analyzer, the polarity score—that is, the Positive, Negative, Neutral, and Compound values—for each daily news item was determined. Higher RMSE values were noted, which meant that the findings did not live up to our expectations. The model was fine-tuned further by include more parameters, such as the exchange rate between USD and INR and gold, Brent, and G-sec.

The model's predictions significantly improved once these parameters were added. The model's RMSE values were similar to those of the previously mentioned LSTM model. For sentiment analysis, the Random Forest Regressor with extra macroparameters was the ultimate solution in this case.

Table 2: RMSE Values aimed at dissimilar models

Models	RMSE
LSTM	39.20
Bidirectional	185.30
Linear Regression	1041.94
Arima	533.75
KNN	1284.16
Prophet	312.57
Random Forest	591.50

International Journal of Research in Engineering and Applied Sciences(IJREAS)
Available online at http://euroasiapub.org/journals.php
Vol. 12 Issue 12, December- 2022
ISSN(O): 2249-3905, ISSN(P) : 2349-6525 | Impact Factor: 8.202 |
Thomson Reuters ID: L-5236-2015

- **Model Evaluation**

  - **LSTM Model**

Regression analysis has been utilized to project future stock prices in order to develop trading strategies based on stock price predictions for this study. Out of all the models we have tried, LSTM has proven to be the most successful in price prediction. A member of the deep learning algorithm family, the LSTM relies on feedback connections within its design. Because it can analyst a whole series of data, it has an advantage over standard neural networks.

In LSTM, data pre-processing is a crucial step. Because most models recommend scaling data, LSTM also has to handle the data in the form of scaling. Given that LSTM bases its single-value prediction on sequences, it is applicable to sequences. As a result, a matrix must be made using the available date-wise train data set. Several model iterations, including the addition of different Dense, Dropout layers, were tried during the model-building process. By comparing mistakes between various runs, hyper parameter adjustment was also done. Although batch normalization was also attempted, the outcomes did not significantly improve. To improve the outcomes, bi-directional LSTM variation was also tried in addition to parameter adjustment.

A sliding historical window of 60 days produced the best results out of all the ranges covered by the model building exercise. The model that performed the best among the many model variations was a two-layer LSTM with 128 and 64 neurons, respectively, followed by two dense layers with 25 and 1 neurons. It is not possible to utilize typical features like accuracy % because this is a regression model. As a result, RMSE was employed as the quantification parameter to assess the effectiveness of the evaluated models.

  - **Random Forest - Sentiment Analysis**

The purpose of this study was to forecast stock prices using sentiment analysis. One of the difficulties with LSTM is that it uses a single parameter to generate models because sentiment analysis could not be done with LSTM. The exercise was divided into two main sections: the

creation of a model for value prediction using sentiment analysis and daily sentiment collecting and analysis.

Analysis additionally, standard libraries were used for preprocessing to enhance the quality of the data. Since there were multiple news items for a given day, the combined news data for the day was obtained by concatenating all of the news items for that particular day. Sentiment polarity was generated using the standard library's Sentiment Intensity Analyzer, yielding four values that corresponded to the input text. It calculates the input text's degree of Positive, Negative, Neutral, and Compound sentiment. For the Sentiment section, these four characteristics are regarded as the independent features. After a few tries with standard regression models, it was determined that the Random Forest Regressor was the best appropriate model for the investigation.

Very subpar outcomes were obtained from the model's preliminary runs that used the close price and the four sentiment features mentioned above. A 20–30% variance was seen in the predicted values. Domain exploration was done to include any other extra features based on the input. Different combinations with external features were tested. Four new features were added to the model: the USD exchange rates, gold prices, Brent price, and GSec yield. These macro-parameters proved to be highly important in increasing the model's prediction accuracy. It is not possible to utilise typical features like accuracy % because this is a regression model. Thus, the success of the evaluated models was assessed using RMSE as the quantifying metric. In the study, the future stock values of four equities for June 28 were predicted using the two models mentioned above, Random Forest and LSTM.

- **Visualizations**

On the closing prices of the four stocks—Reliance, HDFC, TCS, and SBI—LSTM was used. The train, validation, and predict data for the model have been shown in be-low graphs. The train data is shown by the blue line, the validation is shown by the orange line, and the expected close value for the stock is shown by the green line. Of the 3478 total data points, 3305 are used for training and the remaining 5% are used for validation during a 15-year

International Journal of Research in Engineering and Applied Sciences(IJREAS)
Available online at http://euroasiapub.org/journals.php
Vol. 12 Issue 12, December- 2022
ISSN(O): 2249-3905, ISSN(P) : 2349-6525 | Impact Factor: 8.202 |
Thomson Reuters ID: L-5236-2015

period. Reliance, HDFC Bank, TCS, and SBI stocks had RMSE values of 38, 33, 59, and 7, respectively. The error of the LSTM model is substantially less than that of earlier models, such as k-nearest neighbour, ARIMA, and linear regression, among others.

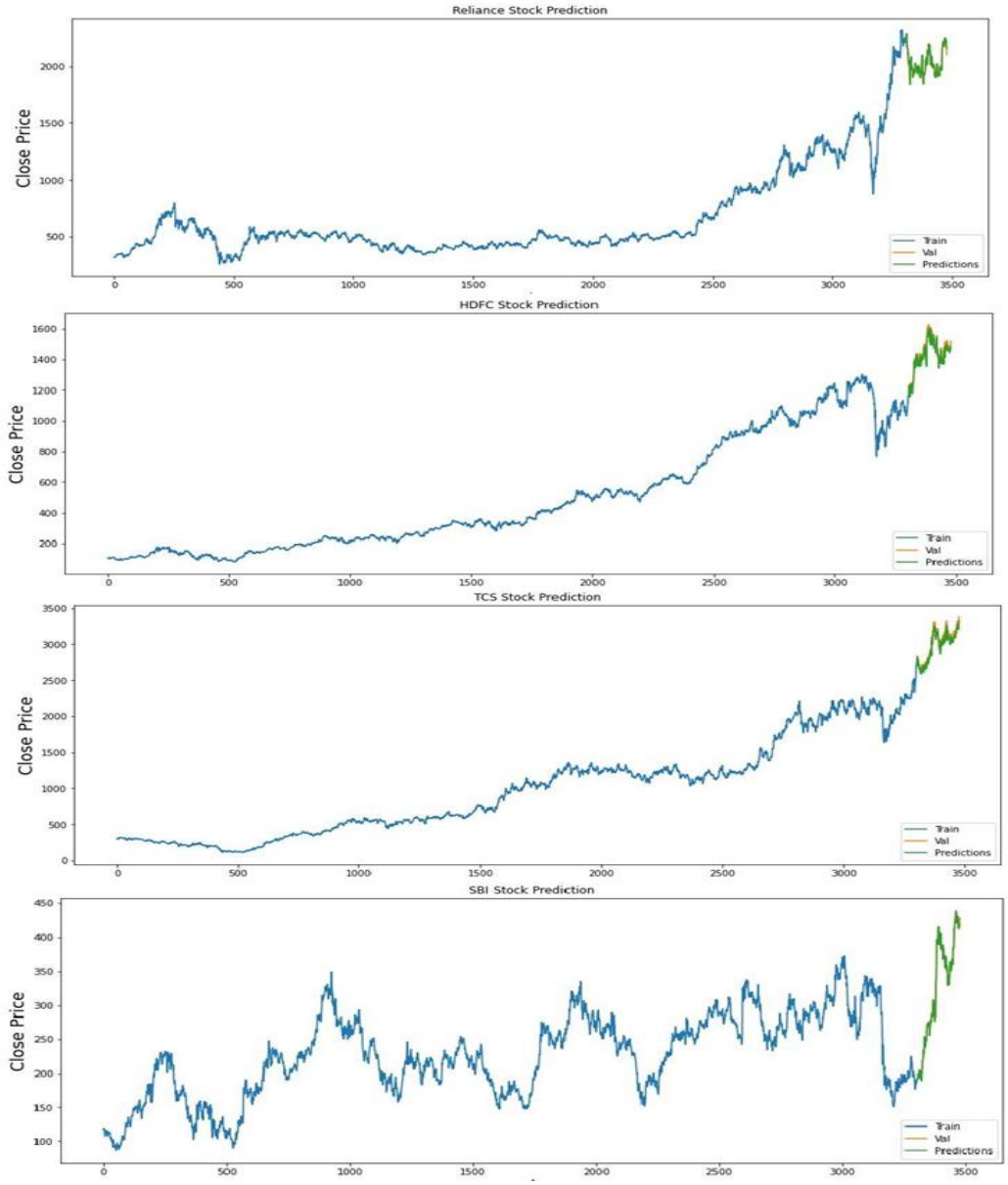

Fig. 4: Stock Estimates as per LSTM

Additionally, an attempt was made to use random forest regression to examine the effects of daily news feelings and outside variables like Gold, G-Sec, Brent, and the INR-USD exchange rate on stock movement. The model's output as a result is shown visually. Although

International Journal of Research in Engineering and Applied Sciences(IJREAS)
Available online at http://euroasiapub.org/journals.php
Vol. 12 Issue 12, December- 2022
ISSN(O): 2249-3905, ISSN(P) : 2349-6525 | Impact Factor: 8.202 |
Thomson Reuters ID: L-5236-2015

LSTM performs better than Random Forest Regression, the Random Forest model does produce better predictions when given more features. The one exception is TCS, whose RMSE value (139) is noticeably more than that of LSTM. Lack of reliable news for sentiment analysis could be one of the causes.

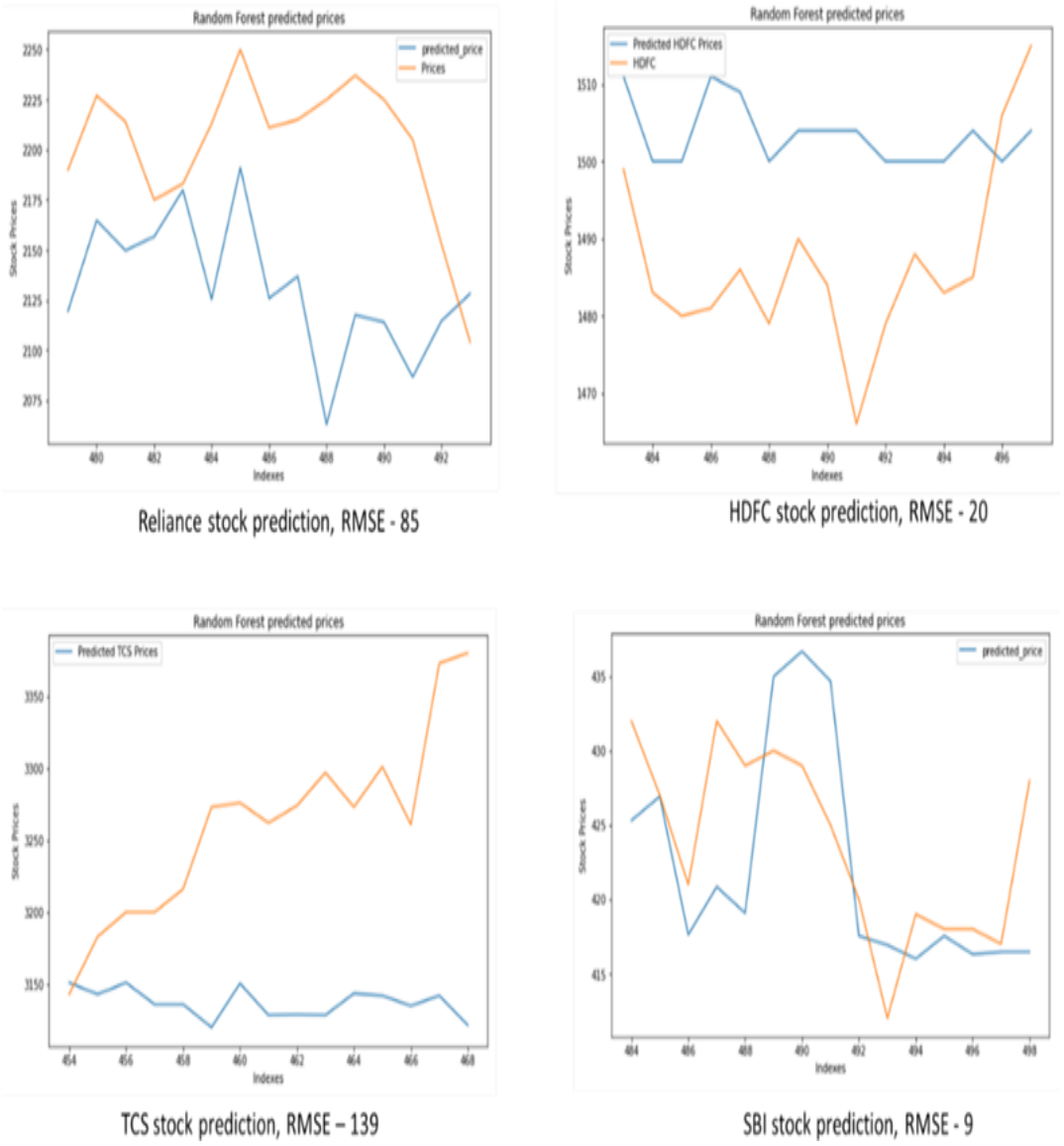

Fig. 5: Stock Calculations as per Random Forest by Sentiment Analysis

- **Conclusion**

The overarching goal of this exercise has been to design trading methods that may facilitate the practical implementation of the generated models. The study was unable to reach those

levels because of the various restrictions and limitations mentioned above.

The next day value was predicted using two different regressors, LSTM and Random Forest, and the assessment metrics were the RMSEs. Since this was a regression exercise, it was not possible to accomplish a prediction with a given degree of confidence. To determine whether the values predicted by our models were reasonable, a basic intuitive analysis of RMSE values was conducted. The following displays the model-achieved results.

Table 3: RMSE standards and Error Ratio aimed at Stock prices

Stocks	LSTM RMSE	LSTM MAPE%	Sentiment Analysis RMSE	Sentiment Analysis MAPE%
RIL	37.20	1.37	86.12	4.42
HDFC Bank	34.15	1.82	21.50	2.26
TCS	60.60	1.61	140.17	4.77
SBI	8.90	1.76	10.66	2.88

The only stock where sentiment research has performed better than LSTM is HDFC Bank. With identical levels of results from Sentiment Analysis and LSTM, SBI is the best-performing stock in both models. As a result, we may state that an approximate fit to explain how this model functions could be defined as having a 95% confidence level. Although no trading strategy was developed throughout the exercise, an effort was made to use the models to predict future price trends rather than simply one day's pricing. The trend predicting findings were unsatisfactory, and in order to have any better results in the future, considerable adjustments may be required.

International Journal of Research in Engineering and Applied Sciences(IJREAS)
Available online at http://euroasiapub.org/journals.php
Vol. 12 Issue 12, December- 2022
ISSN(O): 2249-3905, ISSN(P) : 2349-6525 | Impact Factor: 8.202 |
Thomson Reuters ID: L-5236-2015

References

1. Chen. R and Lazer. M., (2011). Sentiment Analysis of Twitter Feeds for the Prediction of Stock Market Movement, 15.
2. Picasso, A., Merello, S., & Cambria, E. (2019). Technical analysis and sentiment embeddings for market trend prediction. *Expert Systems with Applications*, 60–70.
3. Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The AZFin text system. *ACM Transactions on Information Systems*, 1–19.
4. Tekin, S., & Canakoglu, E. (2018). Prediction of stock returns in Istanbul stock exchange using machine learning methods. 2018 *26th Signal Processing and Communications Applications Conference (SIU)*.
5. Dogan, E., & Kaya, B. (2019). Deep learning based sentiment analysis and text summarization in social networks. *2019 International Artificial Intelligence and Data Processing Symposium (IDAP).*
6. Gallagher, L. A., & Taylor, M. P. (2002). Permanent and temporary components of stock prices: Evidence from assessing macroeconomic shocks. *Southern Economic Journal*, 69(2).
7. Bing, L., Chan, K. C. C., & Ou, C. (2014). Public sentiment analysis in twitter data for prediction of a company's stock price movements. *2014 IEEE 11th International Conference on E- Business Engineering.*
8. Malandri, L., Vercellis, C., & Cambria, E. (2018). Public mood–driven asset allocation: The importance of financial sentiment in portfolio management. *Cognitive Computation*, 1167–1176.
9. Zhang. L., (2013). Sentiment Analysis on Twitter with Stock Price and Significant Keyword Corre- lation, 1-130.
10. Kilimci, Z. H., & Akyokus, S. (2019). The analysis of text categorization represented with word embeddings using homogeneous classifiers. *2019 IEEE International Symposium on INno- vations in Intelligent SysTems and Applications (INISTA).*
11. Li, X., Wu, P., & Wang, W. (2020). Incorporating stock prices and news sentiments for stock market prediction: A case of Hong Kong. Information Processing & Management, 102212.