Predicting The Stock Market Using Contemporary Current Affairs

Group Name: Heavyweights

Shibajee Sarkar(shibajee.prime001@gmail.com)
Anirban Chakraborty(rihanchak@gmail.com)

June 26, 2022

Abstract

One of the most important task today's generation is concerned about is predicting the stock markets, and we all know that the stock market is highly affected by the current affairs. In this study, we suggested a forecasting model for predicting sentiment around stock prices of Bank-Nifty. We map feelings to see if there's a link between news-predicted sentiment and the original stock price, as well as to test the efficient market theory. Finding future stock trends is a difficult endeavour since stock trends are influenced by a variety of factors. Presumably, news items and stock prices are connected. Furthermore, news has the potential to change market patterns. As a result, we set out to investigate this link in-depth and see if stock movements can be forecast using news articles and prior price histories. We have used a pretrained FinBert model for the sentiment analysis and a LSTM for the stock price prediction.

1 Introduction

A popular goal is to develop and/or use a model to sentiment prediction by looking for connections between words and marking them with positive or negative sentiments. There are many opportunities these days to perform sentiment analyses, for example external services that are almost completely ready to use it in a given context where it is needed like TextBlob. In addition, there are options that allow us to create our own models, train them based on our own data. Sentiment analysis with BERT is one of the most powerful tool that we can use, but we can also create a Recurrent Neural Network (RNN) as well.

1.1 What?

In this project we have suggested a forecasting model for closing price of Bank Nifty, where we use economic news headlines as a predictor. We focused on the Indian financial news which are available through news websites. Our primary focus were on the following news websites: (i) Business Standard, (ii) Business Today, (iii) Economic Times and (iv) Indian Express.

1.2 Why?

Several factors influence stock trends, one of which is current affairs, which may have a significant impact on the fluctuation price of the stock as individuals respond to the given news, Nowadays, big influential business tycoons and investors set market trends by publicly criticizing or supporting, daily news articles and platforms serve the purpose of circulating said information to the public and it can influence trading strategies of the stock market. Therefore, it has become necessary to deeply analyze the information to support the investors to make smart trading decisions before making real investments. Our goal is to establish a link between news stories and change in flux of stock movements since there is a delay allying with when news is released and published and how the stock market moves to reflect changes in the value of stocks. Stock market forecasting provides excellent profit opportunities and is a fundamental catalyst for most researchers in this sector. Most researchers use technical or fundamental analysis to predict the market. We are using unquantifiable data in our work, such as news article headlines, to predict stock market trading, which helps us establish a relationship between the two and find trends that can help traders and variety. Here we are solely focusing on the BANK NIFTY index.

1.3 How?

Many people try to interpret and define the different stock market movements in many ways. In this article, we use different tools to the sentiment analysis, especially focussing on the economic news, but in terms of economic news, focussing only on the headlines of economic news. In today's communications and news consumption, the headlines of various articles play an even more important role than before. Now, we use sentiment analysis on these headlines to determine the effects of the headlines to the stock market.

Data is an important pillar of analysis. Primarily the headlines of economic news are needed, what we use for sentiment analysis. We collect the news headlines from popular financial news websites using web-scraping tools like Beautiful Soup and Selenium. Secondly, different stock market data are also needed. There are many possibilities for data collection and analysis, from 'traditional' dictionary-based performed by humans to 'more serious' neural networks that determine the polarity of the headlines of each economic news and label with ap-

propriate emotional polarity. In the case of stock market data, numerous tools are available to obtain stock market data which can be even company-specific which is important to us. In both cases, we work with the most up-to-date data as possible. We are working with data from January 2009 till April, 2022. Both, the headlines of the economic news and stock value data are related to the time period which specified by the news. Thus, the results of the given emotional analysis and the range of stock market data will be appropriate.

The analysis can be separated to the next sections: (i) Collecting headlines of economic news through web-scraping and collecting stock market data. (ii) Then preparing the headlines data and applying FinBert sentiment analysis tool for financial text. (iii) Then the RNN model, LSTM was built and taught using the libraries and capabilities provided by Pytorch.

2 Literature review

In [1], there is a stock market prediction model using combinatoion of LSTM Neural Networks, ARIMA and Sentiment Analysis. This paper presents a methodology of using Long Short-Term Memory (LSTM) cells, a type of RNN, in combination with a time series model, called as Auto Regressive Integrated Moving Average (ARIMA), and a Sentiment Analysis model. The output of these three models are combined in a Feed forward Neural Network for predicting the final value of next day price.

In [2], the paper has primarily implemented a Bert model for sentiment analysis of the news headlines. They also used TextBlob, NLTK – VADER lexicon and RNN for the same. They used different sentiment analysis tools to emotionally analyze and classify different economic news headlines and examine their impact on different stock market value changes even without their full context. Emotions were classified into the usual positive negative and neutral categories. Neutral categories appeared for TextBlob and NLTK-VADER Lexicon tools, but not for Recurrent Neural Network (RNN). The various sentiment analyses results were compared with the result of BERT as a benchmark. The results of the RNN model outperformed the other sentiment analysis tools and gave a result quite close to BERT, emphasizing that there was no neutral emotional value in this case either.

In [3], the researchers have tried to study the effects of current affairs on the Dow Jones Industrial average. They considered a Time Series of DJIA, without news headlines, as their Baseline Model. They showed that using only the stock market values, the performance of the prediction is not promising. To improve this performance, they pre-process the input text, and use it as an input for their LSTM model.

3 Proposed Methodology

Our problem is to make a forecasting model of Bank Nifty closing price using headlines of economic news in India.
The methodology followed in the project is:

- News scraping from economic news websites using Beautiful Soup and Selenium.
- Using FinBert model to bring out the sentiments from the collected news headlines.
- Building an LSTM model to forecast the stock prices.

3.1 Web scraping in our project:

One of the main steps in this project is collection of the news headlines, since we have not used any pre-collected data for news headlines. The data collection is done through web-scraping the news websites of (i) Business Standard, (ii) Business Today, (iii) Economic Times and (iv) Indian Express. There is more discussion about this in the dataset section of Experimental results below.

3.2 FinBert model in our project:

FinBERT is a pre-trained NLP model to analyze sentiment of financial text. It is built by further training the BERT language model in the finance domain, using a large financial corpus and thereby fine-tuning it for financial sentiment classification. Financial PhraseBank by Malo et al. (2014) is used for fine-tuning.
Here is the main article that describes the development of the FinBERT model. The model is trained on 1.8M news headlines and provided a stunning accuracy of 86% which hugely outperformed other models like A plain LSTM model with GloVe embeddings, LSTM model with ELMo embeddings and ULMFit.

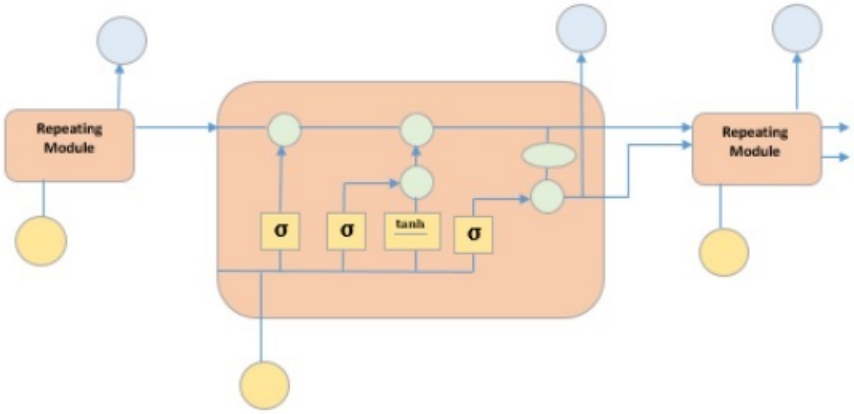
	All data			Data with 100% agreement		
Model	Loss	Accuracy	F1 Score	Loss	Accuracy	F1 Score
1. LSTM	0.81	0.71	0.64	0.57	0.81	0.74
2. LSTM with ELMo	0.72	0.75	0.7	0.50	0.84	0.77
3. ULMFit	0.41	0.83	0.79	0.20	0.93	0.91
4. LPS	-	0.71	0.71	-	0.79	0.80
5. HSC	-	0.71	0.76	-	0.83	0.86
6. FinSSLX	-	-	-	-	0.91	0.88
FinBERT	**0.37**	**0.86**	**0.84**	**0.13**	**0.97**	**0.95**

Experimental results on the Financial PhraseBank dataset

We put the collected headlines through the FinBERT model to get a 1x3 vector of sentiments, each element denoting the proportion of the sentiment present in the sentence, the first vector denoting the proportion of neutral sentiment, second one denoting positive sentiment, the third one denoting the negative sentiment. The sum of the three elements is 1.

3.3 LSTM

LSTM is a special kind of recurrent neural network that is capable of learning long term dependencies in data. This is achieved because the recurring module of the model has a combination of four layers interacting with each other.

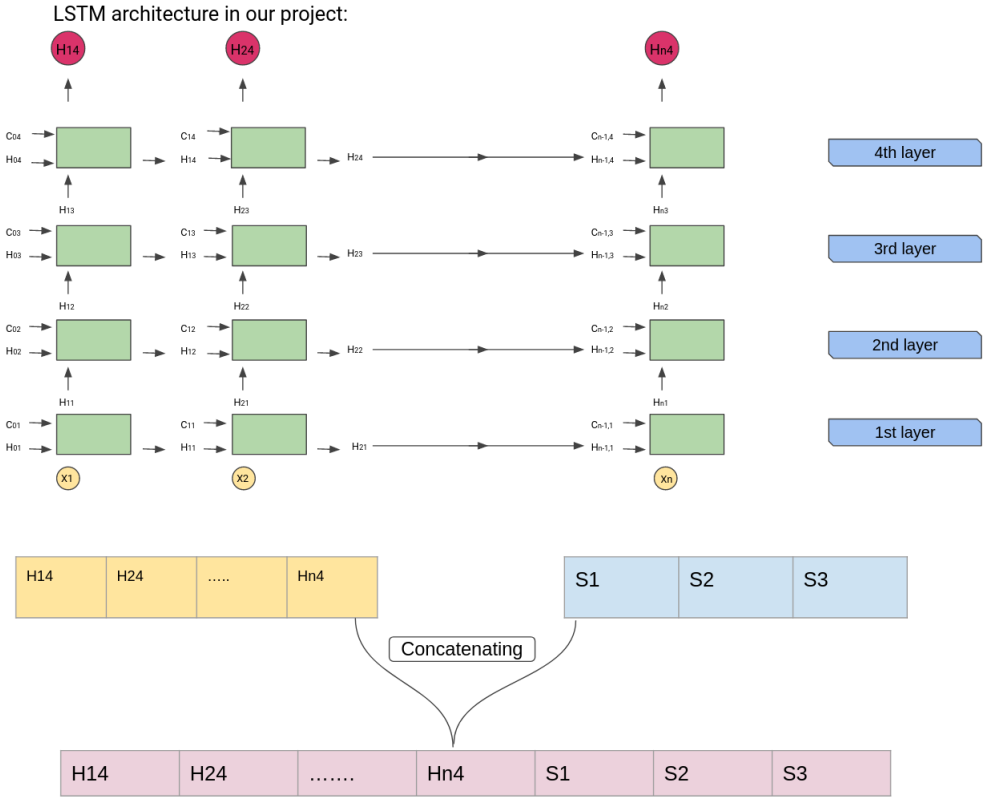


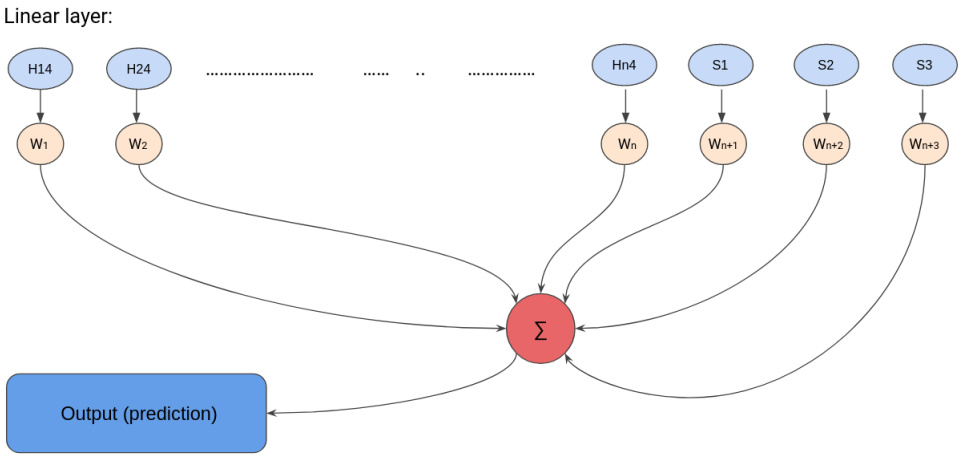
The picture above depicts four neural network layers in yellow boxes, point wise operators in green circles, input in yellow circles and cell state in blue circles. An LSTM module has a cell state and three gates which provides them with the power to selectively learn, unlearn or retain information from each of the units. The cell state in LSTM helps the information to flow through the units without being altered by allowing only a few linear interactions. Each unit has an input, output and a forget gate which can add or remove the information to the cell state. The forget gate decides which information from the previous cell state should be forgotten for which it uses a sigmoid function. The input gate controls the information flow to the current cell state using a point-wise multiplication operation of 'sigmoid' and 'tanh' respectively. Finally, the output gate decides which information should be passed on to the next hidden state.

3.4 LSTM architecture in our model

Our data goes through the LSTM, and to replicate the thinking of a human being, that is seeing the how the news of the previous day affects the next day stock price, the sentiment of the news is concatenated with the last time step hidden state of the LSTM and then we pass that vector through a fully connected linear layer to get the output as our prediction.

We have defined a dataloader to make the input data in a sequential manner of n(we have used n =10 and 5) days. For predicting the closing price of a day, we give the last n days closing price as input data in our neural network. We have taken 10 and 5 hidden layers in our LSTM model with respect to using 10 or 5 days closing price to predict the 11th or 6th day closing price. The last time step hidden state is extracted from the LSTM layers and the sentiment vector is concatenated with the hidden states to pass through a fully connected hidden layer. We have used MSE loss function and ADAM optimizer with learning rate 0.1



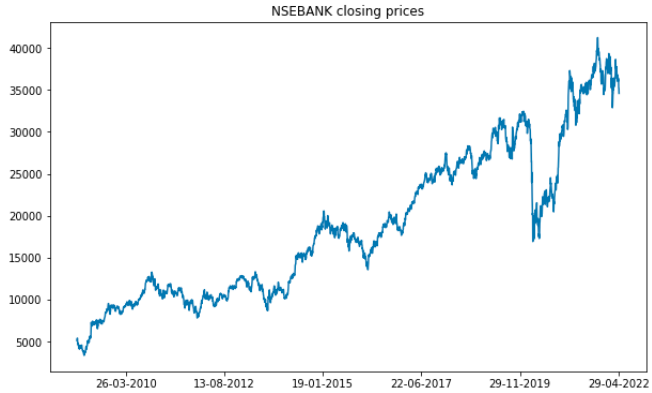


4 Experimental Results

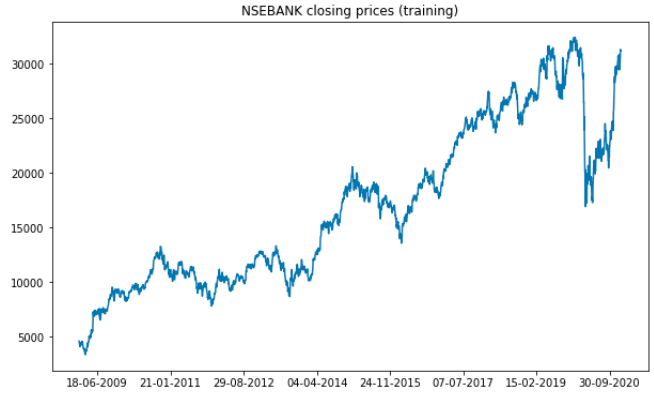
4.1 Datasets

Since we were using a data collected first hand by us, we had to use web scraping tools like Beautiful Soup and Selenium. The websites that we used had either a next page button or a load more button to keep on getting the data. The websites having next page button had to be dealt using Beautiful Soup and the the websites having load more button had to be dealt using selenium. The raw datasets containing headlines corresponding to dates are as follows:
Business Today Indian Express Economic Times Business Standard
We had concatenated the news headlines according to date and used regular expressions to remove unnecessary words (this was done to get the best possible result from FinBERT model). We have collected data from 2009, January till April 2022. This is done so that we get a fairly large dataset. But news data was missing for old news and hence, after finding out the sentiment of the available news, we used a smoothing technique for the missing news. We estimated a missing news sentiment by the mean of the last 7 days sentiment. Also, to use as mush data as possible, we used the mean sentiment of all the available news sentiments for each date. Dealing with the NA values like this leads to some bias.

We also had to collect the data on stock prices of Bank nifty.The collected dataset is as follows:
Bank Nifty stock data from 2009



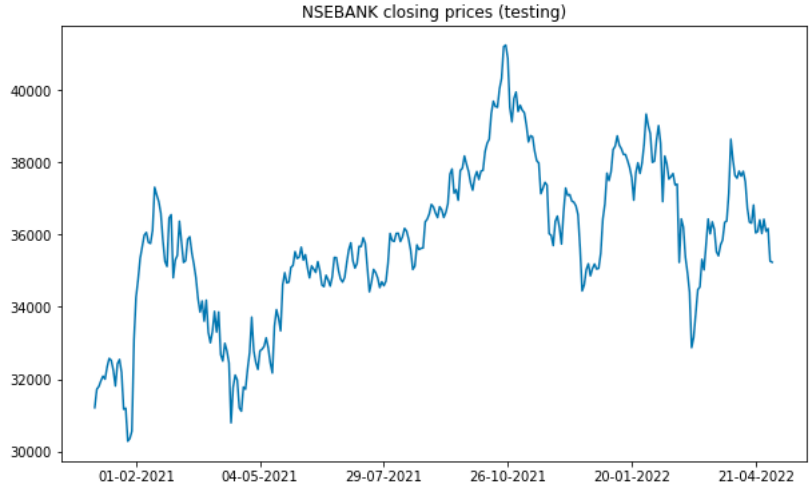
The datasets that we used for training the model:



We first made a stock price prediction using only the past stock price data, without using the news headlines. That is, we predicted the closing price using the starting price, closing price, daily high and daily low. The data used is as follows:
Processed data for prediction without headlines
We used the mean sentiment of the news from the four websites and used that to make our final prediction. As we could not get news for all the dates, we used a smoothing technique. The missing values are replaced by the mean of the last seven days sentiment. The resulting dataset is: Dataset for final analysis

4.2 Experimental Setting

We have made the in five different experimental settings. The models are:
(i) Model 1: Without using headlines, we predict the closing price of the sixth day using five days market data.
(ii) Model 2: Without using headlines, we predict the closing price of the 11th day using 10 days market data.
(iii) Model 3: Using the headline of the Indian Express website only, we predict the 11th day closing price using data from 10 days. This is because we had the most headlines data from the Indian Express.
(iv) Model 4: Using the mean sentiment derived from all the headlines, we predict the 6th day closing price using the data from 5 days.
(v) Model 5: Using the mean sentiment derived from all the headlines, we predict the 11th day closing price using the data from 10 days.
For all the above settings we have used used learning rate=0.01, optimizer = Adam, Loss Function = MSE, number of epochs = 850.



4.3 Experimental Results

Model	No of parameters	training RMSE	testing RMSE
Model1	886	469.53	2556.62
Model2	3171	507.02	1436.05
Model3	3174	353.86	992.68
Model4	889	473.68	1471.77
Model5	3174	423.95	1026.13

From the above table we see the RMSE comparison at different settings of the model. Now we produce below the training and testing performance of the settings through graphs.
It is observable that the all the training sets have great accuracy because the dataset is very large.

Fig1: Model1: training

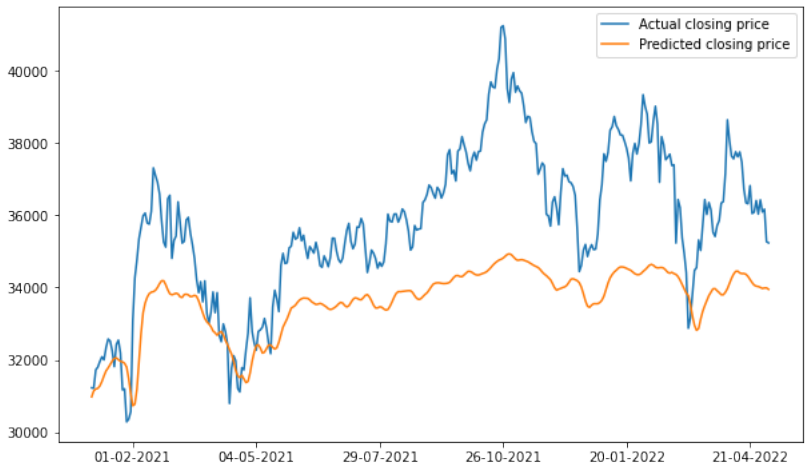


Fig2: Model1: testing

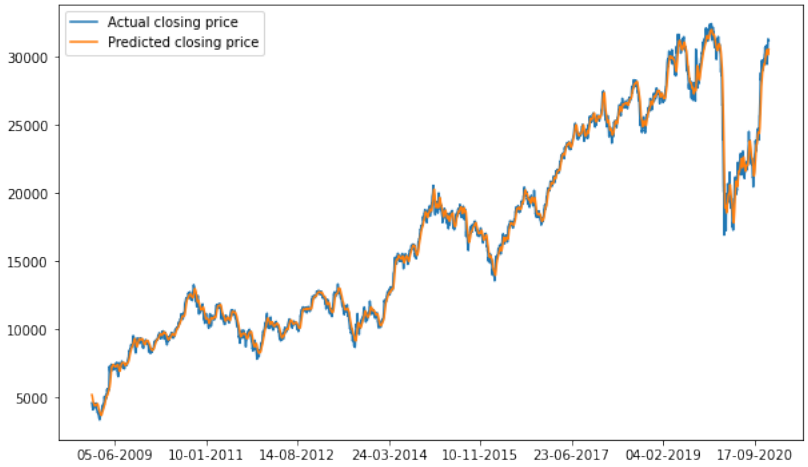


Fig3: Model2: training

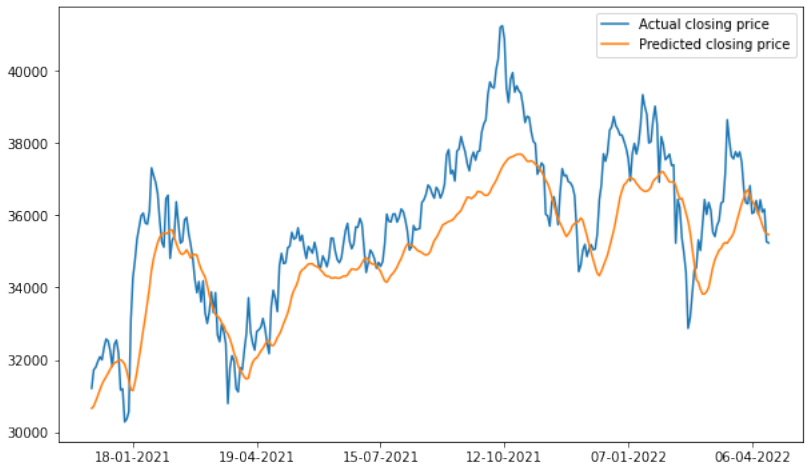


Fig4: Model2: testing

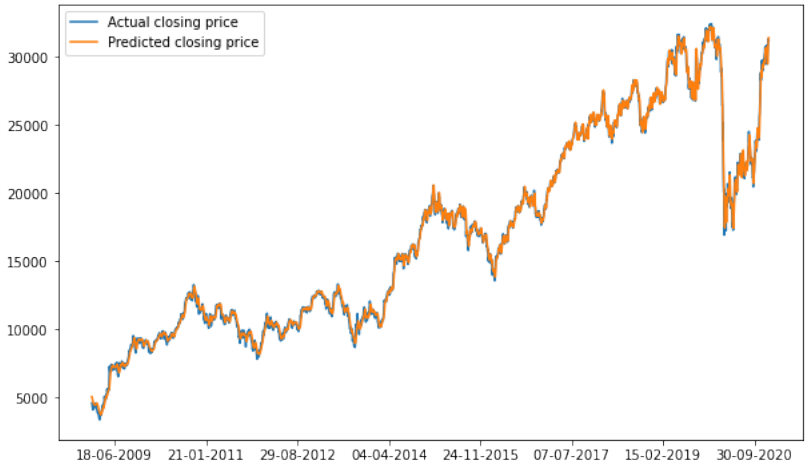


Fig5: Model3: training

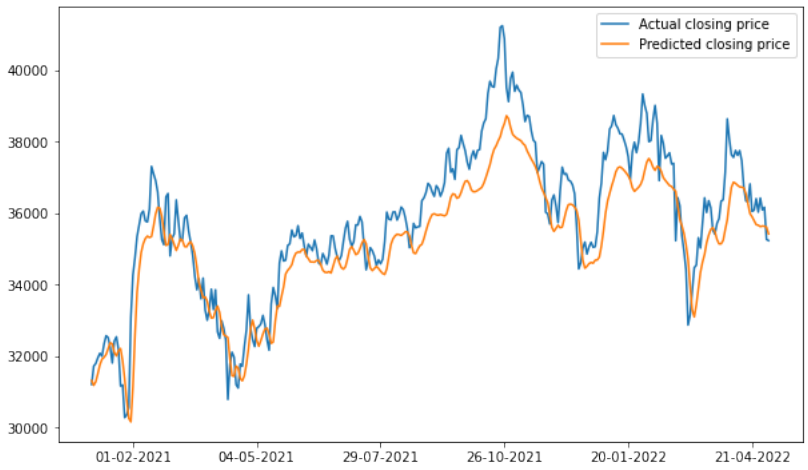


Fig6: Model3: testing

Fig7: Model4: training

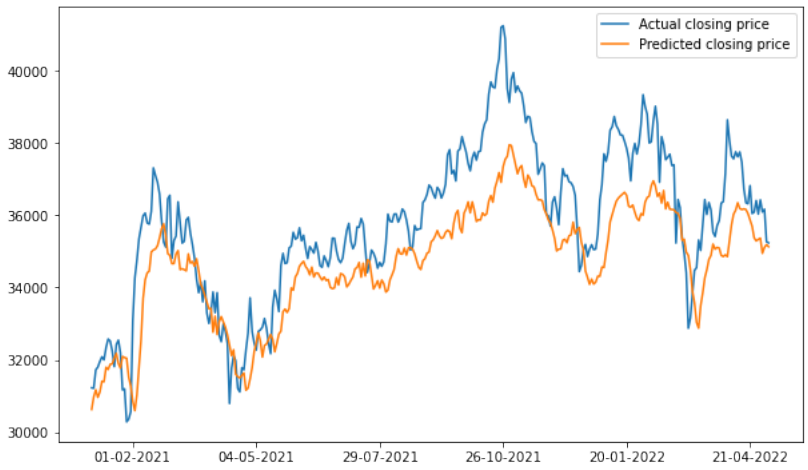


Fig8: Model4: testing

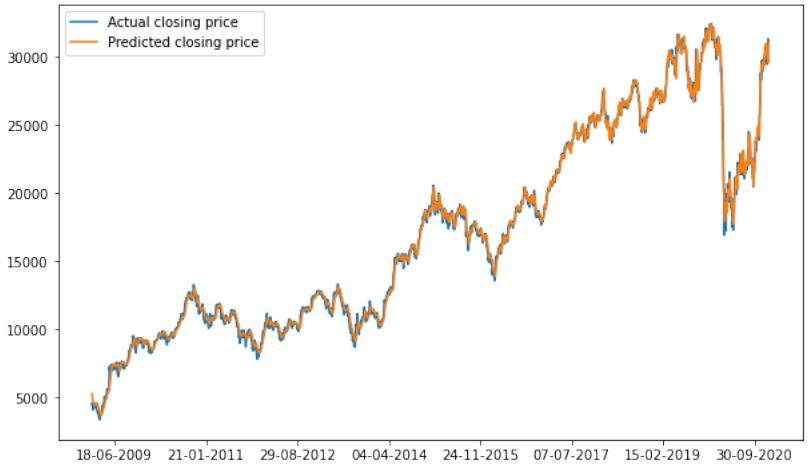


Fig9: Model5: training

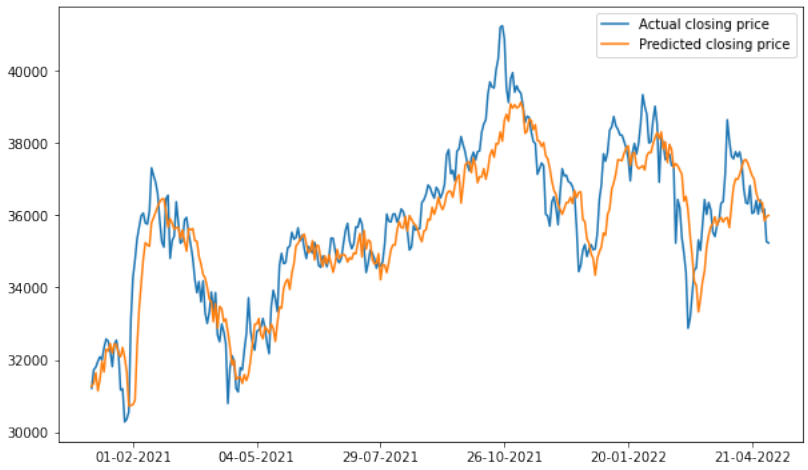


Fig10: Model5: testing

5 Summary

From this project we very clearly visualised that the contemporary current affairs have a great deal of impact on the stock market. The model performed way better when we used the headlines in comparison to when we used the stock price data only.

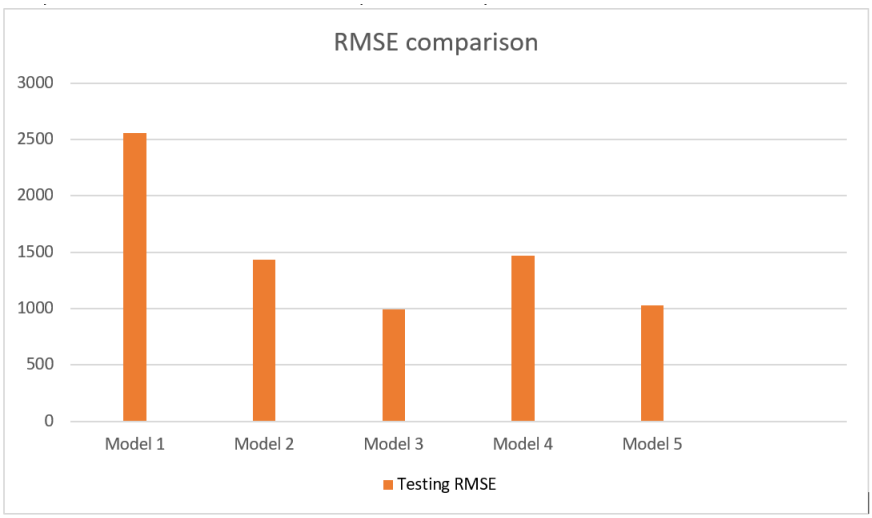


Fig11: RMSE comparison

Considering that the dataset is completely new, with no previous work done on it, we can say that the model has performed considerably well. Our analysis shows that news headlines, which reflect the overall status of economy, politics of the day, society, and so on, can fairly explain a good portion of the hidden variability in the changes in stock market indexes. However, it would be best to study the news headlines alongside a set of other factors that most experts believe to be highly influential on the stock market situation.

There are a number of limitations to this work:

(i) We have only used headlines of the economic news. It would have been better if we could use whole news articles.

(ii) We used only news from India and did not use international news. International news also has the potential to affect the Nifty index in our country.

(iii) We did not get news headlines corresponding to all the timestamps where stock data is available. So, we had to use estimated values for those points, which we did by using a extrapolative smoothing technique. If the data were available, we believe that the prediction would have been better.

Future Work:

(i)We can use better dataset with less missing values and more usable news data for better prediction in the future. Also, we can make more models using different deep learning techniques and time series forecasting models like ARIMA, GARCH etc to have a better overview of the performance of the model.

(ii) Use the news of the days when market is closed. We want to make better predictions by including the news headlines even when market data is not available.

(iii) Statements of business tycoons and influential traders from social media can be analysed and used for better prediction.

(iv) Automating this entire system can lead to prediction of the stock prices at any time of the day. We are looking forward to making an automated system using improved version of this work to predict the stock prices for more frequent timeframe.

6 Reference

[1] Omkar S. Deorukhkar1, Shrutika H. Lokhande2, Vanishree R. Nayak3, Amit A. Chougule4, Stock Price Prediction using combination of LSTM Neural Networks, ARIMA and Sentiment Analysis, International Research Journal of Engineering and Technology (IRJET), Volume: 06 Issue: 03 — Mar 2019.

[2] László Nemes Attila Kiss (2021): Prediction of stock values changes using sentiment analysis of stock news headlines, Journal of Information and Telecommunication, DOI: 10.1080/24751839.2021.1874252.

[3] Ali Hassanzadeh, Razavi Donald, Reza Asadi, Stock Market Prediction using Daily News Headlines, January 2020SSRN Electronic Journal, DOI:10.2139/ssrn.3685530.

[4] Dogu Tan Araci, FinBERT: Financial Sentiment Analysis with Pre-trained Language Models, arXiv:1908.10063v1 [cs.CL] 27 Aug 2019.