Empirical Study and Forecasting Tesla Stock prices using Sentiment analysis and deep learning methods

MSc Research Project
Data Analytics

Jorden Anthon Lopes
Student ID: x19213344

School of Computing
National College of Ireland

Supervisor: Dr.Bharathi Chakravarthi

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Jorden Anthon Lopes
Student ID:	x19213344
Programme:	Data Analytics
Year:	2021
Module:	MSc Research Project
Supervisor:	Dr.Bharathi Chakravarthi
Submission Due Date:	31/01/2022
Project Title:	Empirical Study and Forecasting Tesla Stock prices using Sentiment analysis and deep learning methods
Word Count:	7726
Page Count:	21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	30th January 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Empirical Study and Forecasting Tesla Stock prices using Sentiment analysis and deep learning methods

Jorden Anthon Lopes
x19213344

Abstract

The stock market has become one of the biggest source of income and nowadays many new people have started investing in the stock market. Investment depends on proper timing and in the selection of proper company, and to make this decision making easier this research has been implemented. In this study stock value analysis and prediction has been done for the Tesla Company. For time-series data analysis LSTM, RNN,BI-LSTM, CNN-RNN, CNN-LSTM and customized CNN-BI-LSTM deep learning models have been executed in which the proposed model CNN-BI-LSTM gave better results as compared to other models. The impact of convolution layer and outliers are also being examined, to support this research, sentimental analysis has been carried out on Tweets related to Tesla Company with the help of NLTK VADER libraries. This is a secondary study to understand the impact of tweets on stock prices. The final intention behind the whole study was, before getting involved or investing in any company, investors would get some estimation, so that they could make decisions accordingly also results of implemented experiments could help new researchers to select techniques.

Key words: LSTM ,RNN, BI-LSTM, CNN-RNN, CNN-LSTM, CNN-BI-LSTM

1 Introduction

Is it Possible? To predict future price movements in Share Price. We all know the stock price of the company depicts the investor's expectation of the prospect of the Company. It is directly linked to the future of the investors as well. In a country like India,10 years ago people avoided dabbling into the stock market, but with the advent of technological advances and educational resources available on the Internet have paved the way for many retail traders entering into Stock Market for secondary Income. The stock market is a fundamental part of the Financial Markets. By understanding the need of an hour stock prediction topic has been selected by the researcher (Idrees et al.; 2019).Many scientists have published their research in this sector,still this sector needs an improvement because, some people think that stock market is Casino better to stay away from it, but with proper analysis and study of the previous data, any new investor could make profitable decisions.

By considering the volatile nature of the stock market it is difficult to predict exact pattern (Wen et al.; 2019a), but in the current era, nothing is impossible, as a being new investor, the situation of the newcomer has been taken into consideration and this research topic has been selected. Many investors take suggestions from brokers or financial advisors who always keep eye on any news or leads coming out from targeted companies by taking note of this approach sentiment analysis of tweets published on tweeter have been carried out, the negative tweets may cause bad impact on the company stock data (Liu; 2020). Twitter is one of the major social media platform to appreciate or criticize any organization or any influential person, so to support research on stock prediction, as secondary part sentiment analysis was also carried out.

Tesla is one of the most famous automobile companies in the world. Also in analysis, it has been observed that this stock has many ups and downs which will help to make the solution more robust, so Tesla Company has been selected for stock price prediction. The main intention behind this research is to make new and old investor's life easy in better decision making also by developing robust solutions help new researchers to understand data and models. In this research, the researcher has tried to answer the below question.

RQ1:How effectively stock price prediction could be done by analysing the impact of outliers and convolution layer?

overall three Experiments have been carried out to understand the impact of the convolution layer and outliers in the financial time series data. results provided by all models have been compared with each other and based on evaluation metric MAE , best model has been selected for further prediction process.

The sentiment analysis on 14 days Tweets related to Tesla has been carried out. Where, with the help of positive and negative sentiment scores, the researcher has tried to show actual ups and downs in closing stock price.

The entire process of this study has been described chapter wise.

Chapter 1: This section is the Introduction section which will give an idea about the background of this study, why this topic has been selected for research, the motivation behind this research. **Chapter 2:** This section highlights ideas about related work done in the same or different domain through which the researcher could finalize his methodology and techniques. **Chapter 3:** This section highlights the methodology section which includes subsections like data collection, preprocessing and data mining. **Chapter 4:** This chapter includes design specifications that contain the overall architecture of research, also if the researcher is proposing any new technique then this section includes that information. **Chapter 5:** This is the implementation section which describes sequentially how the entire process is carried out and what are parameters need to be considered while model execution. **Chapter 6:** In this section, researcher evaluates the best performing model based on evaluation metrics and prediction is also done based on the same. This section also has a discussion part where the researcher highlights the pros and cons of applied techniques. The researcher has discussed the shortcomings of the models and has explained to overcome those shortcomings. **Chapter 7:** This is the conclusion section which provides a conclusion based on research findings and in future work, researcher have suggested what advancement could be done in existing research. Technological improvements are making human life easier and will continue to do so. This research might become a small part of that Ocean of Knowledge that could help people to make better decisions in the stock market or it will help any researcher to discover an advanced solution.

2 Related Work

While analyzing stock price data of companies like Tesla, there might be high volatility in the data and to face such volatility and predict better results there is a need of building a robust system that can be developed by implementing versatile models and techniques and for shortlisting those techniques researcher has gone through a plethora of research papers some of them are mentioned below

2.1 Role of Machine Learning and deep learning Algorithms for Predicting Stock Performance

It has been mentioned in this research that a combination of two-directional -two-dimensional principal component analysis and deep learning could improve the accuracy of stock multimedia, as compared to the CNN. This research has been implemented with the help of the Google dataset. By adjusting window sizes and dimensions, model performance got improved. On window size 20 and dimension 10*10 model gave the best results. The model has been compared with Radial Basis Function Neural Network (RBFNN) and Recurrent Neural Network (RNN) and Deep Neural Network (DNN) gave better accuracy than RBFNN and RNN. The RNN model accuracy is very poor and by adding customized layers of other models or proper data preprocessing this issue could have been resolved. Also proposed model is not performing well in terms of total return and Root Mean Square Error (RMSE) as compared with the RBFNN model (Singh and Srivastava; 2017).

The data set consists of 1721 NSE listed company and minutes wise stock price has been selected. Companies from the IT sectors and Pharmaceutical sectors have been selected. For the shortlisting best model, the RMSE matrix has been evaluated. Execution of RNN, LSTM and CNN deep learning models have been done. From the studies, it has been seen that deep learning models could understand the empirical relationship and predict stock prices. CNN model is capable to find out trends in stock prices so it has been considered as the best model, remaining models have been used in the analysis of other time-series data, it has been mentioned in research that changes happening in stock data are not always following the same pattern. It gets change according to companies and sectors and this research gives more concentration on CNN models as they depend on current information due to which it provides better results (Selvin et al.; 2017).

In this research work, two datasets have been examined. The first datasets were obtained from tick data. The change in the price from trade to trade is getting stored as a tick. The stock price at the start of every 15 minutes has been scrapped from tick data which is used as the second dataset. The variations of ANN models have been implemented which are Levenberg-Marquardt (LM), Bayesian Regularization(BR) and Scaled Conjugate Gradient(SCG).To understand the number of iterations(epochs) and mean square error, a performance plot has been analyzed. To understand the network performance regression plots have been analyzed. The Tick by tick dataset provides better results as compared to 15 minutes dataset. In less time SCG algorithm provides better results as compared to LM and BR models. As suggested in future work in this implementation RNN or LSTM could have been used and to support this study sentimental analysis could be helpful (Selvamuthu et al.; 2019).

By considering the Indian economy the concept of time series analysis and forecasting have been implemented. The author has nicely explained the need of stock price prediction to save economy. To understand the market volatility ARIMA model has been implemented, data which has been used in the execution is publically available and the forecasted values have been compared with actual values which show discrepancy of 5 per cent. For validation ADF test and the L-jung box tests have been implemented. It has been observed that ARIMA model is performing quite good but to understand current market situation this model needs to be taken into advanced level (Idrees et al.; 2019).

In this research novel approach has been introduced for time series trend prediction, this approach works based on reconstructing time series with the help of high order structures like motifs. The data set is extracted from the Yahoo finance website. The patterns in the newly developed sequences were examined by convolution neural networks which provided relatable information about ups and downs. The results have been compared with existing sequential models such as Recurrent Neural Networks(RNN) and it has been observed that the proposed model performed well in terms of computational complexity. The performance of the model has been tested on a real financial time series dataset and it successfully captured trends in the stocks (Wen et al.; 2019b).

In this research to predict next day's stock closing price, five companies from different sectors have been selected and their stocks prices have been analyzed. With the help of stock features like Open, Close, High, Low new features have been developed and supplied to the model for prediction. The artificial neural network (ANN) and random forest (RF) algorithm have been used for prediction. The principle of the decision tree has been used which is automatically get executed in RF algorithm, due to huge size data it consists noise in data and because of that tress grows in wrong directions. The model aims to minimize the forecasting error. Two models have been compared with each other, and the best model has been selected based on Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE) and Mean Bias Error (MBE). According to the matrix, ANN gives better results as compared to RF (Vijh et al.; 2020).

For this research stock prices of Tata Consumer Product company has been selected from year 2013 to 2018 and the data set carries features like Open, Close, Low and High prices. Sequence to Sequence modelling has been used to predict the output based on provided input. In this activity encoder-decoder method has been used in which the encoder converts input into foxed length vector and decoder predict the results based on encoded input. Each OLHC price is taken as a subtask and by using a sliding window input sequence has been created. The window is selected in such a way that loss would be maintained as little as possible and the output of each task is provided to the BI-LSTM model. The Seq2Seq model and multitask model have been executed which gave RMSE values 3.98 and 7.87, these models have been compared with other existing models as well (Mootha et al.; 2020).

In this research, the CNN-BI-LSTM-AM method has been used to predict the next day closing price. This model is the combination of convolution layer, Bi-directional LSTM and Attention mechanism. An attention mechanism is used to improve the accuracy of the model. this model has been compared with other models like CNN, RNN and CNN-LSTM in which CNN-BI-LSTM-AM provided better model accuracy (Lu et al.; 2020).

From the above literature work, it has been observed that deep learning models like RNN, LSTM and BI-LSTM are widely used in stock prediction problems and by adding convolution layer their performance could increase, also CNN-LSTM and CNN-RNN are also used in many research but CNN-BI-LSTM is hardly used in stock prediction or dataset like yahoo finance. So custom CNN-BI-LSTM model would be proposed in this research.

2.2 Tweet analysis for stock price prediction

In this research sentiment analysis is assumed as text classification problems and a Random forest algorithm has been used to classify sentiments in the Indonesian language. For sentiment Bag of words (BOW) library has been used which has weighting methods such as binary TF, raw TF, logarithmic TF and TF.IDF. With Random forest experiments have given a better performance (Fauzi; 2018).

In this research L1 logistic regression, L2 logistic regression and cat boost algorithm have been used in which cat boost algorithm performed better than the remaining two algorithms in the movie review database. The movie review data has been downloaded and with the help of the TF-IDF algorithm data has been converted into vector data. And on this vector data model building process has been carried out (Yang; 2020).

In this research data is extracted from Twitter, Stocktwits and the Yahoo finance website. The sentiment score has been calculated with the help of TextBlob and VADER libraries and for prediction support vector machine(SVM) and Logistic regression algorithms have been used, For model evaluation Accuracy, F-score and Area under curve metrics have been used. SVM performed well in terms of F score and AUC. The researcher has mentioned, in such an imbalanced dataset the combination of SVM with VEDER provides the best result in terms of F score and AUC (Nousi and Tjortjis; 2021).

in this part of the research stock prediction is not a priority task, in this part researcher is just depicting the impact of sentiment on stock data, so NLTK VADER lexicon utility is used for sentiment analysis, this is smart utility to find sentiment score also it outperform text blob (Bonta and Janardhan; 2019) and prediction is just supporting activity done in this section so by considering data set size Random forest, Cat-Boost and customize ANN model will be executed for model prediction.

3 Methodology

Post studying multiple research papers, the methodology for this research paper has been determined. Some of the well-known methodologies are Knowledge Discovery in Database (KDD), CRISP-DM and SEMMA, according to the research topic, researcher selects type of methodology and this project is go under KDD methodology. This methodology carries 6 steps which start from data collection to final result discussion. All the steps have been discussed in the below points. Figure 1 depicts the KDD process of this research.

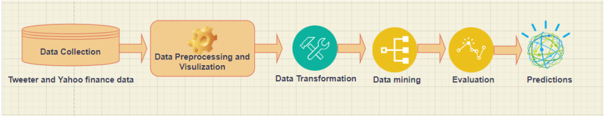


Figure 1: KDD Methodology

3.1 Data collection

Data collection is one of the basic and most important steps in any research project implementation because the outcomes of the projects are highly dependent on the quality and credibility of the data. Some data sets require more pre-processing as raw data could not be passed to the model for prediction, so it takes so much effort. Hence, by considering all such points data has been selected. This project is divided into two parts and for this implementation, two data sets have been used.

The first dataset is collected from the Yahoo finance website which is an open-source website, from this website anyone can download stocks data with the required date range. Consequently, for this study data of Tesla automobile company has been downloaded from 29th of June 2010 to 3rd of February 2020 and the dataset has 7 features that are Date, Open, High, Low, Close, Adjusted Close, Volume. The low and high value represents the minimum and maximum stock price for the provided period. While open and close value represents the stocks starting and ending price for the same period. Volume is the total amount of stock trading which has been done for the period.

The second dataset is downloaded from Twitter, the second part is the implementation of a sentimental analysis of Tweets related to Tesla company. Twitter has developed one API through which developer account holder could download tweets related to any burning topic, with the help of the tweepy library of python, the individual could access that API, bypassing API key, API secret key, Access token, and secret access token key which will be provided with developer account access, Tweets have been downloaded. for this dataset 14 days, tweets related to Tesla have been downloaded which have a count of around 325177 and the features of the dataset are Keyword, Content, ConversationID, Date, Retweetcount, and Tweet Url. Some of the features are not useful for the process so they have been removed in pre-processing part. The Content and Date are the significant features of this raw dataset.

3.2 Data Pre-Processing

Two datasets have been selected for this project implementation , both the datasets were in raw format. To make them more readable or understandable both datasets have gone through pre-processing part.

The first data set which has been downloaded from the Yahoo finance website is first imported into Python notebook and null values have been checked and if exist then they have been removed. with the help of a distribution plot, data distribution has been analyzed. With the help of a correlation matrix highly correlated features have been analyzed in this research Open, Low, High, and Close these columns are highly correlated with each other but they have not been removed as this dataset has limited features and all the columns are part of the prediction. With the help of a box plot outliers have been analyzed and removed with the help of Z score due to variability in historical data. The date column has been converted into Month, Year, and Day format.

The second data set went through many pre-processing steps as removing unnecessary columns and changing column names. Null values have been identified and removed. Some special characters which were like noise in data have been removed from the data. The date column has been converted into date-time format also in the Twitter data set for a single day couple of thousand tweets got downloaded so day-wise all tweets were merged as a single statement so the day-wise data set would have single sentences.

3.3 Transformation

Many changes have been done to the features of the dataset to make the process understandable and executable. In the Yahoo finance dataset as well as in the Twitter dataset Date column has been transformed to respective Day, Month, and Year format. For data visualization and understanding suitable graphs have been added. Results are also showcased in graphical format. Unnecessary columns are removed from both datasets to make the process easier. Through Twitter around 325177 tweets were downloaded for 14 days so day-wise tweets are merged into single sentences so eventually, the Twitter dataset would have 14 rows with tweets and dates as features. To build a better model Yahoo finance data has been merged with the Twitter dataset and a new feature has been added into the Twitter dataset which would have close value for those 14 days records. Newly developed datasets would have a date, tweets, and Close value column. For sentiment analysis and to find sentiment score more 4 columns have been added into the new dataset that is Compound, Negative, Neutral, and Positive. After some execution, it has been observed that the Neutral column has not any significance in the model building so the Neutral column has been removed from the dataset.

3.4 Data Mining

The main reason behind this research is to understand trends in tesla stock prices and accordingly stock prediction would be done. By considering the volatile nature of the stock market there is a need for a robust system that would show its versatile behaviour to manage ups and downs. The deep learning models LSTM, BI-LSTM and RNN have already shown better results in research like (Selvin et al.; 2017) ,(Mootha et al.; 2020) ,(Rather et al.; 2015) Thus by considering them as base models, further implementation has been done by adding convolution layer, therefore CNN-RNN and CNN-LSTM are also widely used models in stock prediction problems which have given satisfactory results in (Zulqarnain et al.; 2020),(Lu et al.; 2020) so the same convolution layer has been added in Bi-LSTM model which created CNN-BI-LSTM model which has been rarely used in stock predictions.

This model exploits the characteristics of the convolution layer and BI-LSTM layer which would show high computational power than other shortlisted models. Such advanced implementation has been done in (Lu et al.; 2021) where an advanced attention mechanism(AM) has been added. From this research proposed model has been selected and performance of the model has been improved by parameter tuning.

In the second part of this project, to support research done with time series model implementation, sentimental analysis has been done, the main intention of this part is to show the impact of sentiments on stock prices. For sentiment analysis, the VADER lexicon feature has been used as this library outperformed Text blob in (Bonta and Janardhan; 2019) , because of the small dataset size basic machine learning models like Random Forest and CatBoost algorithms have been selected for prediction based on sentiment score provided by VADER lexical library.

3.5 Evaluation Metric

Evaluation metric plays a very important role in every implementation as based on the evaluation metric the capability and efficiency of the model could be determined. Evaluation metrics can not be selected randomly there has to be a relation between the data and the type of problem the researcher addressing. Thus this is a regression type of problem in which day-wise different stock prices would be predicted by the model and in such regression problems root mean square error(RMSE) and mean absolute error(MAE) are widely used evaluation metrics (Liu and Chen; 2020), so to select the best evaluation metric from RMSE and MAE, deep analysis of data has been done. This is a stock prediction problem in which, for some reason suddenly stock prices go high for a couple of days, and again prices are coming to the normal range so those high values would act like outliers. In this research separate study has been done on this issue. so the formulas of RMSE and MAE has been mentioned below

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$

$$\text{MAE} = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n}$$

Formula of RMSE and MAE

RMSE is the standard deviation of errors of the predictions, which are also known as residuals, in Figure 2 residuals are picturized.

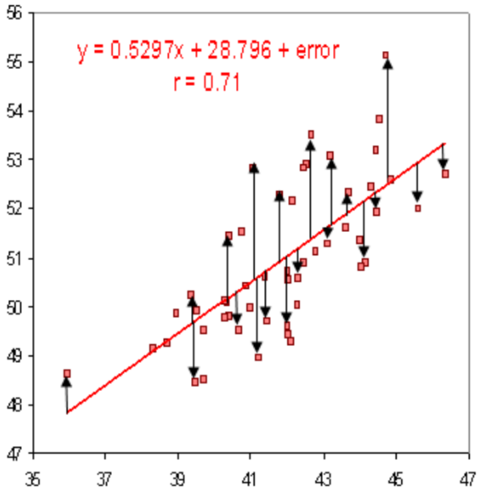


Figure 2: RMSE: Data points location from Line of best Fit

Residual are normally how far data points are situated from the line of best fit where the model could give best results and in-stock prediction data, some data points go far from the line of best fit due to uncertain growth or fall in stock and as per formula of RMSE the square of difference could provide some difficult results and make execution more complex, so mean absolute error metric has been selected which is also a difference between predicted and actual results and equally divided by sample size. which made execution easier as well as made the process more understandable (Willmott and Matsuura; 2005).

4 Design Specification

This section includes a description of architecture implemented in the research project also briefly describes if any novel model or technique was executed according to a dataset or problem statement.

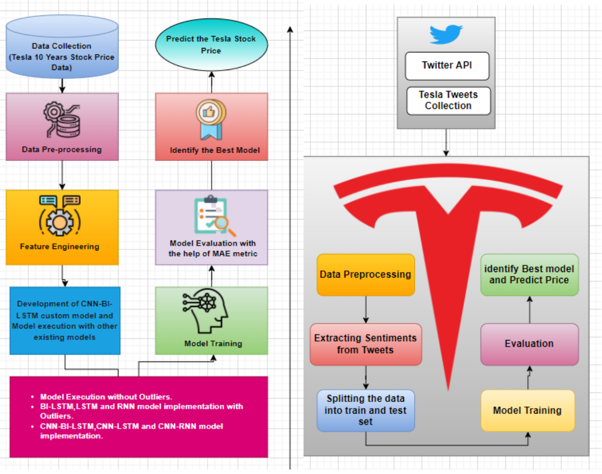

Data Collection (Tesla 10 Years Stock Price Data)
Data Pre-processing
Feature Engineering
Development of CNN-BI-LSTM custom model and Model execution with other existing models
Model Execution without Outliers.
BI-LSTM,LSTM and RNN model implementation with Outliers.
CNN-BI-LSTM,CNN-LSTM and CNN-RNN model implementation.
Model Training
Model Evaluation with the help of MAE metric
Identify the Best Model
Predict the Tesla Stock Price
Twitter API
Tesla Tweets Collection
Data Preprocessing
Extracting Sentiments from Tweets
Splitting the data into train and test set
Model Training
Evaluation
identify Best model and Predict Price

Figure 3: Project Architecture Diagram

As described in Figure 3 this research is a combination of two implementations, the first part includes time series model implementation whereas the second part includes sentimental analysis for stock price prediction. In the first part, deep analysis and visualization of data have been carried out. LSTM, BI-LSTM, and RNN such kind complex algorithms have been implemented on the Yahoo finance dataset. for stock prediction Convolution LSTM and Convolution RNN algorithms are already used and these algorithms have already provided great results in previous research but still stock forecasting is unpredictable and to reduce that uncertainty bi-directional-LSTM model has been used which has high computational power than simple LSTM algorithms so in this research convolution layer has been added in BI-directional LSTM so that this model will use the properties of both CNN and BI-LSTM algorithm. The convolution layer has been installed on a filter size of 256, kernel size of 4 with RELU as an activation function.

The Lambda layer has been added in the model and because of that model could predict all four Open, Low, Close, High values simultaneously. Otherwise, to predict different features, the model has to be executed that many times.It has been mentioned in the methodology section that MAE has been selected as an evaluation metric but to find MAE Stochastic Gradient Descent (SGD)optimizer has been used. The main reason behind using this optimizer is, it will easily get fit into memory because in every iteration single training sample will get processed by the network, this property makes it computationally fast. For large datasets, it works outstanding as it updates parameters frequently and due to frequent updates there would be oscillations in steps taken toward minima of loss functions which would eventually help to get out of the case of local minima of the loss function. Also through SGD learning rates and momentum could be set, through which model execution speed and learning rate could be managed, this functionality is hard to perform in other optimizers.

In this implementation use of the Huber loss function is being done .By considering stock data it would have many outliers and this function is less sensitive to outliers, when loss goes high it changes the equation from quadratic to linear, which can be explained in laymen terms as RMSE to MAE, so this function Is the combination of RMSE and MAE, so this algorithm handles outliers and situation of local minima (Sen et al.; 2021).

In the second part, sentimental analysis implementation has been carried out on Twitter data, post executing plethora of pre-processing steps, Date wise tweets have been merged into single-single statements. Yahoo finance data set is being merged with tweeter data set and a new dataset has been generated. CatBoost, Random Forest, and customized artificial neural network algorithms have been implemented to predict with the help of a newly developed dataset.

5 Implementation

In previous research like (Selvin et al.; 2017) ,the implementation of LSTM and RNN algorithms has been already done. In the case of stock price data, these algorithms have already proved their efficiency but stock data is very uncertain so sometimes best-performing models also fail, so continuous improvement is required in this field.

5.1 Time series data analysis and Predictions

The first dataset is downloaded from the yahoo finance website. post-implementation of data pre-processing, the readable and executable dataset has been divided into train and test datasets and proportion is kept as 90%: 10%. The first BI-LSTM model has been implemented which has two bi-directional LSTM layers with 64 unit sizes also both the BI directional layers have been followed by a dropout layer(0.2) to avoid overfitting of the model. It has been mentioned in (Sunny et al.; 2020) that, with proper parameters tunning, model performance could be improved. So in the final model 2 dense layers have been added with 128 and 64 units, to pass neurons to the next level, the RELU activation function has been used. Customized lambda layer has been added into all models to provide 4 output instead of 1 output as these models will provide High, Low, Open, and Close values in prediction.

The second implemented model is LSTM in which two LSTM layers with a unit size of 64 and 16 have been added and LSTM layers are followed by the Dropout layer (0.2). In base model 2 dense layers have been added with unit size 43 and 64 respectively and the same like previous model RELU activation function has been used.

In the third model, the RNN algorithm has been implemented, in which 2 simple RNN layers have been added with unit size 50 and 16. In the base model, the linear activation function has been used. In the final model 2 dense layer have unit size 43 and 64 being added. In the base model, the Relu activation function has been used.

The convolution neural network model has characteristics to focus on the most obvious feature so it brings improvement in the feature engineering process. So to improve model performance convolution layer has been added in all three models. Convolution LSTM and Convolution RNN models are already implemented in stock price prediction which has given better results in the past same convolution layer has been added in the BI-LSTM model which would have 256 filters with kernel size 4 and has Relu activation function. This custom CNN-BI-LSTM model has been developed to exploit properties of convolution neural network and bI-directional LSTM model. In CNN-LSTM and CNN-RNN same convolution layer has been added with filter sizes 32 and 256. The rest of all configurations were kept the same for comparison.

The below configuration would remain the same for all models. Input has been passed to the model In the sequence of 64 data points and models are predicting 4 outputs so they will accept data in 4 arrays so the (4,64) size of frame has been developed and passed to the models. The Keras learning rate schedular library has been used and with the help of SGD optimizer learning rate has been calculated. the momentum is set as 0.9 to speed up the process, the momentum can be kept between 0 to 1 . Huber loss function has been installed to find the learning rate. The models have been trained on 100 epochs to find the learning rates. Post getting the lowest learning rate ratio full model would be trained on 1300 epochs. MAE and loss in the training process have been calculated for all models also predictions have been done on test data. Graphically difference between actual and predicted values has been shown, also based on the MAE evaluation matrix best model has been suggested.

5.2 Twitter data sentiment analysis

The second dataset is downloaded from the Twitter API, in pre-processing unnecessary columns have been removed from the dataset. The Close value column of the yahoo finance dataset has been merged with the Twitter dataset. NLTK libraries have been imported for sentimental analysis. VADER lexicon utility has been used for processing. Compound, Negative, Positive, and Neutral columns have been added to the dataset and through libraries, respective values have been calculated. tweeter data has been downloaded for 14 days so the new dataset would have 14 rows and Date, close, compared, negative, and positive columns. The data set has been divided into train and test parts where 9 rows are allocated in the training dataset and 5 rows are allocated in the test dataset.

The first Random forest algorithm has been implemented and predictions have been done. The second cat boost algorithm has been implemented with 500 iterations, a Mean Absolute Error has been calculated for the training dataset also the difference between actual and predicted values have been shown graphically. Deep learning libraries have been imported and an Artificial Neural Network(ANN) model has been implemented, this is a custom model which has 7 dense layers and a dropout layer has been added with 0.2 units in each dense layer. Linear and Relu activation function has been used and the best optimizer Adam has been attached where loss has been calculated with the help of a Mean Absolute Error. The model has been executed on 200 epochs. Predictions have been done and mean absolute error has been calculated to compare with other models to decide the best model.

6 Evaluation

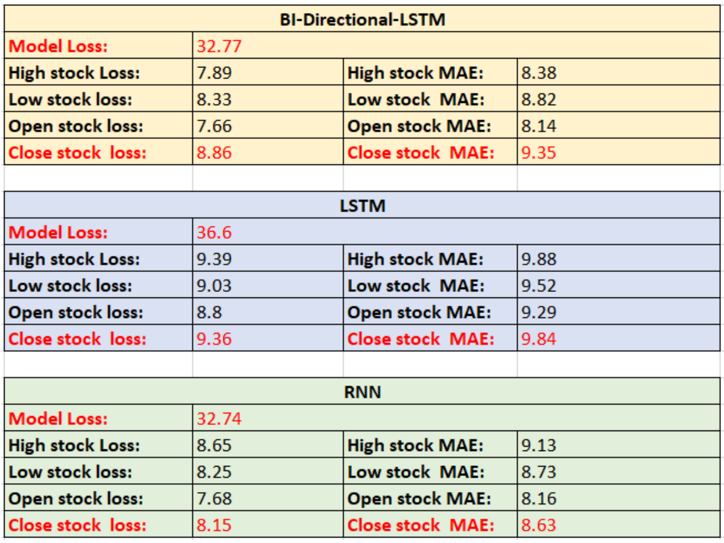
In this section, model performance has been evaluated based on the evaluation metrics. The significance of the Mean Absolute Error evaluation metrics has been explained in the methodology section.

6.1 Time series models implementations

In the first part of this research, time series model implementation has been done, to understand the efficiency and significance of the data and models, also adopted techniques are working properly or not three experiments have been implemented, results of these experiments will help in developing solution which may be useful In academic and industrial purpose.

6.1.1 Experiment 1: Implementation of simple BI-LSTM,LSTM and RNN models

To understand the impacts of the convolution layer the initial models have been compared with each other. The LSTM, BI-LSTM, and RNN models have shown better performance in previous research so these models have been implemented in the yahoo finance dataset. Total loss of models, loss in each feature, and MAE of each feature have been showcased in below table. In this experiment, RNN has shown better performance than LSTM and BI-LSTM models in case of loss in the entire model loss and MAE of Close Stock.without adding convolution layer simple models are compared with each other and results are sequentially depicted in below table.



BI-Directional-LSTM			
Model Loss:	32.77		
High stock Loss:	7.89	High stock MAE:	8.38
Low stock loss:	8.33	Low stock MAE:	8.82
Open stock loss:	7.66	Open stock MAE:	8.14
Close stock loss:	8.86	Close stock MAE:	9.35

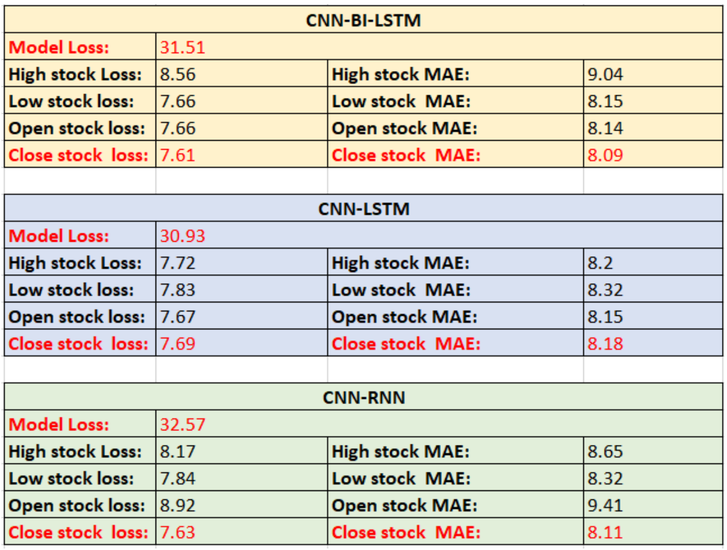
LSTM			
Model Loss:	36.6		
High stock Loss:	9.39	High stock MAE:	9.88
Low stock loss:	9.03	Low stock MAE:	9.52
Open stock loss:	8.8	Open stock MAE:	9.29
Close stock loss:	9.36	Close stock MAE:	9.84

RNN			
Model Loss:	32.74		
High stock Loss:	8.65	High stock MAE:	9.13
Low stock loss:	8.25	Low stock MAE:	8.73
Open stock loss:	7.68	Open stock MAE:	8.16
Close stock loss:	8.15	Close stock MAE:	8.63

Simple BI-LSTM, LSTM and RNN model outputs

6.1.2 Experiment 2: Adding CNN layer In BI-LSTM,LSTM and RNN model

The implementation of CNN-LSTM and CNN-RNN has been done in previous research, so a custom model has been developed in which a convolution layer has been added in the BI-LSTM model. Convolution layer has characteristics of feature extraction which boost the model performance, so post adding convolution layer, three models which are CNN-BI-LSTM, CNN-LSTM, and CNN-RNN have been executed in the same environment. This experiment intends to understand the impact of the convolution layer on simple models. below table depicts the outcome of the newly implemented models. As expected after including the convolution layer into models, the loss in the overall models got decreased and the close stock loss and close stock MAE of CNN-BI-LSTM model are less than CNN-LSTM and CNN-RNN, means developed customized model (CNN-BI-LSTM) will perform better than remaining models while predicting close stock values. In this experiment, the close stock loss was 7.61 and the close stock MAE was 8.09 of CNN-BI-LSTM model, the main intention behind this experiment was to understand the impact of convolution layer , according to data results may get change but in this experiment loss in all the three models got decrease so ultimately performance of the models increased.



CNN-BI-LSTM			
Model Loss:	31.51		
High stock Loss:	8.56	High stock MAE:	9.04
Low stock loss:	7.66	Low stock MAE:	8.15
Open stock loss:	7.66	Open stock MAE:	8.14
Close stock loss:	7.61	Close stock MAE:	8.09

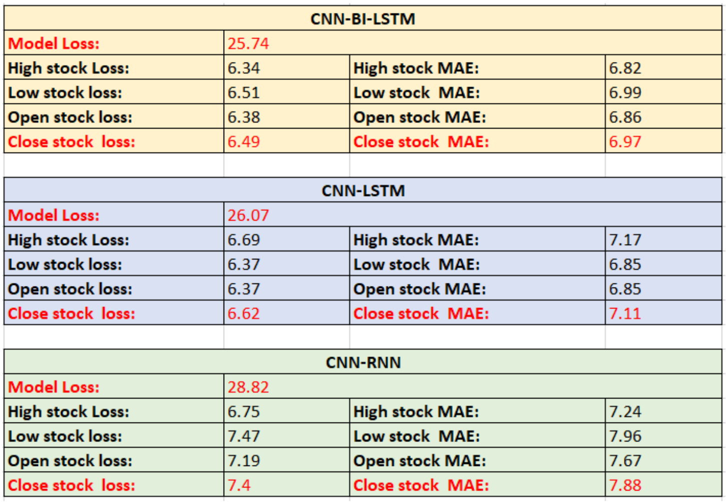
CNN-LSTM			
Model Loss:	30.93		
High stock Loss:	7.72	High stock MAE:	8.2
Low stock loss:	7.83	Low stock MAE:	8.32
Open stock loss:	7.67	Open stock MAE:	8.15
Close stock loss:	7.69	Close stock MAE:	8.18

CNN-RNN			
Model Loss:	32.57		
High stock Loss:	8.17	High stock MAE:	8.65
Low stock loss:	7.84	Low stock MAE:	8.32
Open stock loss:	8.92	Open stock MAE:	9.41
Close stock loss:	7.63	Close stock MAE:	8.11

CNN-BI-LSTM, CNN-LSTM and CNN-RNN model outputs

6.1.3 Experiment 3: Understand the impact of Outliers

In data pre-processing, some outliers have been found which might act as noise in the process. As in-stock price dataset outliers indicate sudden growth or fall in the stock prices which does not usually occur, and therefore these outliers do not help in making patterns. As a result, these are likely to pose difficulty in stock prediction. So in this experiment outliers have been removed from the dataset and all customize models have been executed again results are mentioned in below table



CNN-BI-LSTM			
Model Loss:	25.74		
High stock Loss:	6.34	High stock MAE:	6.82
Low stock loss:	6.51	Low stock MAE:	6.99
Open stock loss:	6.38	Open stock MAE:	6.86
Close stock loss:	6.49	Close stock MAE:	6.97

CNN-LSTM			
Model Loss:	26.07		
High stock Loss:	6.69	High stock MAE:	7.17
Low stock loss:	6.37	Low stock MAE:	6.85
Open stock loss:	6.37	Open stock MAE:	6.85
Close stock loss:	6.62	Close stock MAE:	7.11

CNN-RNN			
Model Loss:	28.82		
High stock Loss:	6.75	High stock MAE:	7.24
Low stock loss:	7.47	Low stock MAE:	7.96
Open stock loss:	7.19	Open stock MAE:	7.67
Close stock loss:	7.4	Close stock MAE:	7.88

Model results post removing outliers

After the removal of outliers, it has been observed that the Loss and MAE of all models have gone down, which indicates performance got increased especially in the CNN-BI-LSTM model which showed less model training loss than CNN-LSTM and CNN-RNN models. The close stock loss and close stock MAE value are also smaller than the remaining two models. Hence, it has been proved that if outliers are handled properly in stock prediction data, then model performance could be increased. The best prediction model is CNN-BI-LSTM which has an overall model training loss of 25.74, the close stock loss is 6.49 and close stock MAE is 6.97.

6.2 Twitter data Sentimental Analysis

Post data has been divided into train and test datasets. Model implementation has been done. The first model implemented was a Random Forest regressor with 10000 estimators, the predictions have been done. And mean absolute error has been calculated based on actual and predicted values. The value of MAE is 42.84 so the graphical comparison between actual and predicted values of 5 predictions have been shown in Figure 4

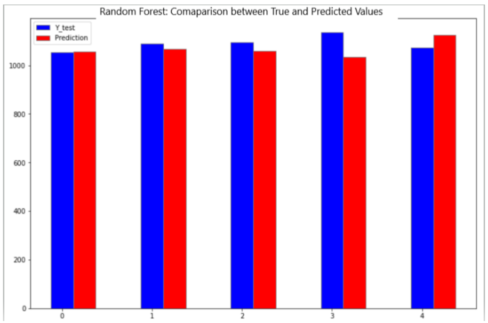


Figure 4: Random forest model prediction comparison

The CatBoost classifiers libraries have been imported and executed in 500 iterations. The model has been fit for the training dataset. The MAE of the model is 37.6 which is less than the Random Forest algorithm. The third implementation is done with the help of the ANN algorithm which gave a mean absolute error of 85.84 which is so high than the Random Forest and CatBoost algorithm and therefore, in this case, ANN is failing. The graphical comparison of 5 prediction of CatBoost algorithm has been shown in Figure 5

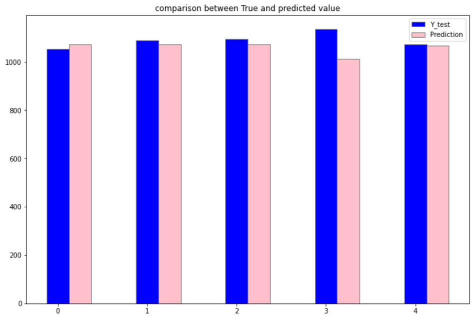


Figure 5: Catboost algorithm model prediction comparison

6.3 Discussion

In this section overall results of the experiments would be discussed. All experiments have been executed on a single system that provided the above results, these results may vary from system to system as the researcher has already executed on different

2 systems and he recorded slight different outputs, CNN-BI-LSTM is a comparatively complex algorithm as compared to CNN-RNN and CNN-LSTM so according to data and system it may work better or poor, but it has been observed that acquired results are close to each other, there is no huge gap in the results. Therefore, these results could be marked as final results.

6.3.1 Discussion on Time series implementation

This research is divided into two parts in the first part researcher is concentrating on an empirical study of Tesla stocks with the help of deep learning models. The models like LSTM, RNN, and BI-LSTM are selected as a base model because they have already performed well in previous research which has been mentioned in the literature review section. In this research entire methodology and techniques have been adopted by performing a deep literature review. So, to understand the quality of the dataset and impact of the convolution layer basic models have been executed and they have given satisfactory results in which the RNN model has performed better than LSTM and BI-LSTM in case of model loss, Close stock loss, and Close stock MAE value.

The lambda layer has been added in all models so that it could give 4 outputs, thus like earlier there is no need to execute models again and again for prediction of different features. In single model execution, predictions of HIGH, LOW, OPEN, and CLOSE stock values could be done. This research is mainly focused on Close value.

It has been observed in previous research that CNN-LSTM and CNN-RNN models have been already implemented in such stock price prediction problems. And CNN-BILSTM-AM model implementation is also done which have already given effective results, so in second experiment convolution layer has been added into all basic models through which customize model has been created known as CNN-BI-LSTM and post-execution, models have given surprising results total learning loss value has been decreased for all models and MAE value also got reduced in all models. Through this experiment, it could be concluded that because of the convolution layer model performance got increased. CNN-LSTM model has overall less learning loss than CNN-BI-LSTM and CNN-RNN model but the value of close stock loss and close stock MAE is less in the CNN-BI-LSTM model. In the pre-processing section, it has been observed that the Yahoo finance dataset carries some outliers. Experiments have been done on all models with outliers and without outliers. It completely depends on the number of outliers. Essentially, outliers come in the stock dataset due to sudden growth or fall in stock values which are not usual activities. Hence, these outliers do not help in pattern formation which is not good in the prediction process still in some datasets number of outliers is high so they play a significant role so they cannot be removed from the dataset.

In this implementation number of outliers was 57 so those outliers were removed, and experimentation has been done which gives outstanding results. There was more fall in model learning loss values and MAE values of all features. In this case, the customized model CNN-BI-LSTM performed better than other models the model loss value was observed at 25.74 and close stock loss and MAE value were 6.49 and 6.97.

By considering the condition of Outliers, the MAE evaluation metric has been selected for evaluation as MAE is more robust to the outliers than MSE. There might be a situation where outliers could not be removed due to their significance, so to overcome this problem use of the Huber loss function has been done which is a combination of MSE and MAE (Chen et al.; 2017). Depending on the delta value which is also known as a hyperparameter, if outliers exist in data then the delta value will be high so the process will go for the equation of MAE which is a linear equation otherwise it will go with the equation of MSE which is a quadratic equation. The formula of Huber loss function has been shown below.

$$L_\delta(y, f(x)) = \begin{cases} \frac{1}{2}(y - f(x))^2 & \text{for} |y - f(x)| \le \delta, \\ \delta\,(|y - f(x)| - \frac{1}{2}\delta), & \text{otherwise.} \end{cases}$$

Formula of Huber loss function

This function is also used to reduce local minima. If an outlier does not exist, then the process will go for the MSE equation which has a quadratic equation and through which the system will get global minima and the system could penalize the error by squaring it. Also to find loss concerning learning rate SGD optimizers have been used Figure 6. This shows the selection of maximum learning rate with respect to the lowest loss value which is 1e-6 (this value gets changed according to data and noise in the data)

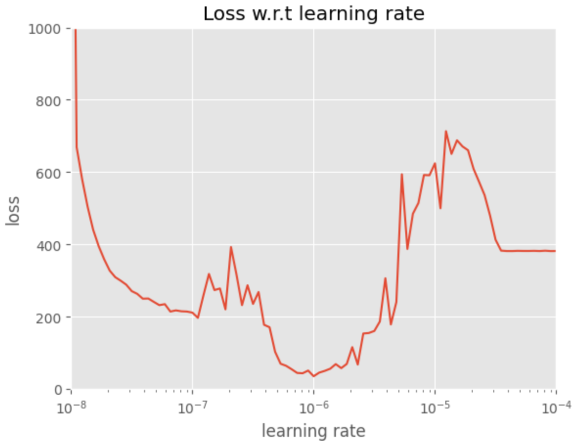


Figure 6: Loss with respect to learning rate

this value has been passed to SGD with a momentum of 0.9 so that process of learning would be fast. Post execution training loss has been showcased in Figure 7.

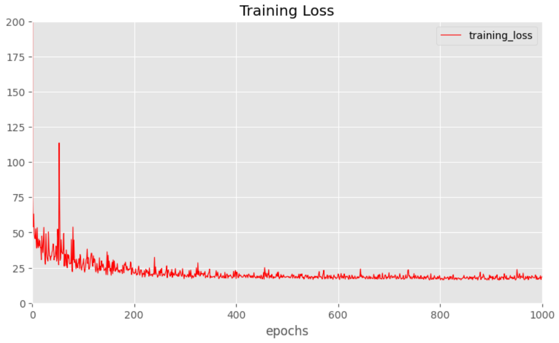


Figure 7: Training Loss

By removal of outliers, there was some drop in loss of MAE values of all features, but the difference was not so high and through this, it could be said that outliers are handled by the system and in the future, if a large number of outliers exists in the dataset then this implementation could handle that situation.

In this complete execution, CNN-BI-LSTM performing slightly better than other models so the graph of prediction and actual values have been shown in Figure 8.

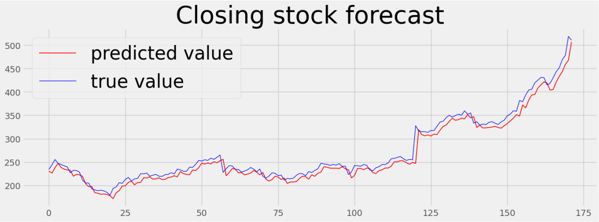


Figure 8: Prediction VS. True value graph

But these results may change slightly from system to system but there would not big difference in the results and it does not mean that CNN-BI-LSTM is performing better means the remaining model failed, if the evaluation value would be checked then it could be observed that all models MAE and Loss values are close to each other. In this research, the researcher has introduced two sub-questions.

To answer the first question researcher has implemented CNN-BILSTM, CNN-LSTM, and CNN-RNN models, and the performance of the models have been compared with the help of MAE evaluation metric, Figure 9 depicts that CNN-BILSTM models have less MAE so in this research stock prediction would be done by this model and practically predictions have been successfully done by the system.

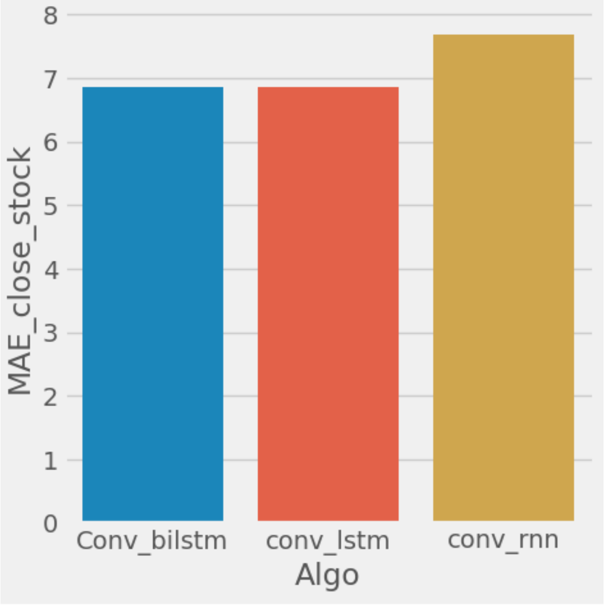


Figure 9: Model comparison w.r.t MAE

By implementing experiments 2 researchers have proved that by adding convolution layer model performance was improved, also by removing outliers model gave slightly better performance functioning of Huber loss function explained which handles the outliers.

6.3.2 Discussion on sentimental analysis implementation

A huge amount of research has been done in stock price prediction with time series model implementation, so to support the study which has been implemented in the first part, sentimental analysis implementation has been done. The data has been downloaded from Twitter API. Due to hardware limitations and time constraints for this research, 14 days of tweets(7/11/2021 – 20/11/2021) have been downloaded. Which have a count of 438896. All data pre-processing steps have been implemented where unnecessary columns have been removed.

The main intention behind this implementation is to find sentiment value and accordingly show the impact of those sentiments on stock value. For that Yahoo finance data set has been merged with the tweeter dataset. Therefore, the final dataset would have a Date, Tweets, and Close value as a column. Some close values were not present so those values were replaced with a mean value of the close value column. with NLTK VADER sentiment analyser, sentiment score has been calculated which would add Positive, Negative, and Compound column values in the dataset, by replacing tweets column. The new dataset would have 14 rows and 5 features.

Post dividing into train and test datasets Random Forest, CatBoost, and Artificial neural network algorithms have been implemented. The mean absolute error of the RF model is 42.84 and the cat boost model is 37.6 but the MAE of the ANN model is 85.84, it could be said that the ANN model is failing to predict close stock values. The dataset size may be one of the reasons behind the failure of this model . Figure 10 depicts the model comparison based on the MAE value of each model. .

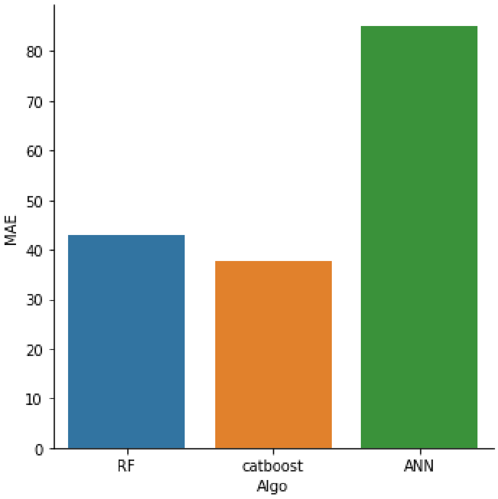
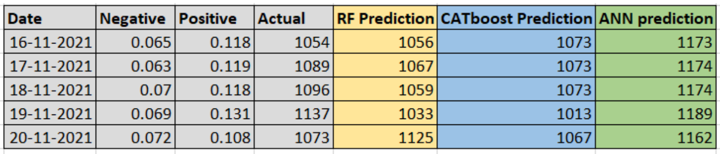


Figure 10: Model comparison w.r.t MAE

To support this research , the researcher has compared the sentiment score calculated from the VADER library to the actual stock closing value. And it has been observed that there be can relation framed between sentiment positive and negative value and stock closing price. below table describes according to positive score there is growth observed in actual stock value. The remaining three columns are predictions done by the machine learning models.The goal of this implementation was to understand relation between sentiment score and actual stock value so that in future more efficiently stock prediction could have been implemented. as per the outcomes which are mentioned in below table it could be seen that through this experiment relation has been successfully proved.



Date	Negative	Positive	Actual	RF Prediction	CATboost Prediction	ANN prediction
16-11-2021	0.065	0.118	1054	1056	1073	1173
17-11-2021	0.063	0.119	1089	1067	1073	1174
18-11-2021	0.07	0.118	1096	1059	1073	1174
19-11-2021	0.069	0.131	1137	1033	1013	1189
20-11-2021	0.072	0.108	1073	1125	1067	1162

Combined outcomes of sentiment analysis study

Due to hardware limitations and time constraints, only 14 days of data has been downloaded by the researcher but with proper hardware support if more day's data is acquired then sentiments and stock value relation could be more clearly explained. The process of calculating sentiment intensity takes around 4 hours to complete, if advanced pre-processing techniques are applied it could make a sentence easier to understand by which execution time will get reduced. everyday a couple of thousands of tweets get published and, in this implementation, all tweets related to tesla have been downloaded. By applying advanced filtration, only highly effective tweets could be shortlisted which would make execution easier.

7 Conclusion and Future Work

This research is focused on analysing and forecasting the stock performance of Tesla automobile company. With the smart decision investors could earn good returns in a short period. In order to assist Investors in decision making, this research have been implemented .In this study LSTM, RNN and Bi-LSTM deep learning algorithms have been compared with complex deep learning algorithms like CNN-LSTM, CNN-RNN and CNN-Bi-LSTM. The impact of the convolution layer and outliers have been examined in this research and relative workaround also been discussed which would be somehow helpful for research community. Comparatively, CNN-BI-LSTM performed better but results are close to each other so sequence may change, to support this study sentimental analysis on tweets related to Tesla have been carried out and researcher has tried to show according to positive and negative sentiment score, ups and downs in stock values. Random Forest, CatBoost and ANN algorithm have been executed and compared with each other and in this CatBoost provided better prediction than RF and ANN algorithm.

With a deep study on Hyperparameter tuning, existing model performance could be increased, also on the existing dataset, new customized model could be executed to get better predictions. With high tech hardware configuration, large volume of Twitter (Tweets) data could be downloaded, also with advanced filtration, quality tweets can be shortlisted, so that advanced deep learning and machine learning models can be implemented for better predictions.

8 Acknowledgment

First I would like to thank my supervisor Dr.Bharathi Chakravarthi for supporting me and guiding me throughout this process and special thanks to God and my family for making me capable of this study.

References

Bonta, V. and Janardhan, N. K. N. (2019). A comprehensive study on lexicon based approaches for sentiment analysis, *Asian Journal of Computer Science and Technology* **8**(S2): 1–6.

Chen, C., Yan, C., Zhao, N., Guo, B. and Liu, G. (2017). A robust algorithm of support vector regression with a trimmed huber loss function in the primal, *Soft Computing* **21**(18): 5235–5243.

Fauzi, M. A. (2018). Random forest approach fo sentiment analysis in indonesian, *Indonesian Journal of Electrical Engineering and Computer Science* **12**(1): 46–50.

Idrees, S. M., Alam, M. A. and Agarwal, P. (2019). A prediction approach for stock market volatility based on time series data, *IEEE Access* **7**: 17287–17298.

Liu, C.-L. and Chen, Q.-H. (2020). Metric-based semi-supervised regression, *IEEE Access* **8**: 30001–30011.

Liu, X. (2020). Analyzing the impact of user-generated content on b2b firms' stock performance: Big data analysis with machine learning methods, *Industrial marketing management* **86**: 30–39.

Lu, W., Li, J., Li, Y., Sun, A. and Wang, J. (2020). A cnn-lstm-based model to forecast stock prices, *Complexity* **2020**.

Lu, W., Li, J., Wang, J. and Qin, L. (2021). A cnn-bilstm-am method for stock price prediction, *Neural Computing and Applications* **33**(10): 4741–4753.

Mootha, S., Sridhar, S., Seetharaman, R. and Chitrakala, S. (2020). Stock price prediction using bi-directional lstm based sequence to sequence modeling and multitask learning, *2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, IEEE, pp. 0078–0086.

Nousi, C. and Tjortjis, C. (2021). A methodology for stock movement prediction using sentiment analysis on twitter and stocktwits data, *2021 6th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM)*, IEEE, pp. 1–7.

Rather, A. M., Agarwal, A. and Sastry, V. (2015). Recurrent neural network and a hybrid model for prediction of stock returns, *Expert Systems with Applications* **42**(6): 3234–3241.

Selvamuthu, D., Kumar, V. and Mishra, A. (2019). Indian stock market prediction using artificial neural networks on tick data, *Financial Innovation* **5**(1): 1–12.

Selvin, S., Vinayakumar, R., Gopalakrishnan, E., Menon, V. K. and Soman, K. (2017). Stock price prediction using lstm, rnn and cnn-sliding window model, *2017 international conference on advances in computing, communications and informatics (icacci)*, IEEE, pp. 1643–1647.

Sen, J., Dutta, A. and Mehtab, S. (2021). Profitability analysis in stock investment using an lstm-based deep learning model, *2021 2nd International Conference for Emerging Technology (INCET)*, IEEE, pp. 1–9.

Singh, R. and Srivastava, S. (2017). Stock prediction using deep learning, *Multimedia Tools and Applications* **76**(18): 18569–18584.

Sunny, M. A. I., Maswood, M. M. S. and Alharbi, A. G. (2020). Deep learning-based stock price prediction using lstm and bi-directional lstm model, *2020 2nd Novel Intelligent and Leading Emerging Sciences Conference (NILES)*, IEEE, pp. 87–92.

Vijh, M., Chandola, D., Tikkiwal, V. A. and Kumar, A. (2020). Stock closing price prediction using machine learning techniques, *Procedia Computer Science* **167**: 599–606. International Conference on Computational Intelligence and Data Science.
URL: *https://www.sciencedirect.com/science/article/pii/S1877050920307924*

Wen, M., Li, P., Zhang, L. and Chen, Y. (2019a). Stock market trend prediction using high-order information of time series, *IEEE Access* **7**: 28299–28308.

Wen, M., Li, P., Zhang, L. and Chen, Y. (2019b). Stock market trend prediction using high-order information of time series, *Ieee Access* **7**: 28299–28308.

Willmott, C. J. and Matsuura, K. (2005). Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance, *Climate research* **30**(1): 79–82.

Yang, Z. (2020). Sentiment analysis of movie reviews based on machine learning, *2020 2nd International Workshop on Artificial Intelligence and Education*, pp. 1–4.

Zulqarnain, M., Ghazali, R., Ghouse, M. G., Hassim, Y. M. M. and Javid, I. (2020). Predicting financial prices of stock market using recurrent convolutional neural networks, *International Journal of Intelligent Systems and Applications* **12**(6): 21–32.