Predicting Stock Movements Using News Headlines And News Articles

[1]Dr. Gresha Bhatia, [2]Deepak Tejwani, [3]Kuldeep Singh, [4]Rohit Vinod, [5]Shubham Shinde
[1]Deputy Head of Department of Computer Engineering, Vivekanand Education Society's Institute of Technology, Chembur, Maharashtra, India.

,[3,4,5,]Students, Department of Computer Engineering, Vivekanand Education Society's Institute of Technology, Chembur, Maharashtra, India.

Abstract

The main objective of the system is to analyse the future value of a certain stock of a particular company using the sentiment analysis and to predict whether a particular stock will go up that is whether it will increase or it will go down which means it will decrease on the basis of certain news headline, also detection of fake news and OCR was implemented for providing the user as an option for entering the news headline or a news article, the data we used was DJIA news headlines dataset and five different machine learning algorithms were used – Random Forest classifier, Naïve Bayes, Decision Tree, Logistic Regression and Support Vector Machine(SVM).

***Keywords**: Sentiment analysis, OCR, Naive Bayes, DJIA, Decision Tree, Logistic Regression, Support Vector Machine*

1. INTRODUCTION

Predicting stock prices and the status of stock market is a quite strenuous task in itself. Today stock prices of any company so not depend upon the financial factors of the company but also on the various other factors such as socio-economic factors and especially in this century the movement of stock prices are no more only linked with the current economic situation of the country rather the stock prices of the particular day are also directly or indirectly depends on the company related news, natural calamities as well as the political events. The motive of the research is to build a machine learning model which will predict whether the stock price of a company will go up or will go down and the model also predicts the exact stock prices for the next day and the day after based on the today's news headlines of the company. We have taken Dow historical stock dataset of the years 2008-2016 which consists of the date, news headline, stock movement labelled as '1' for increment and '0' for decrement. The OCR model was also integrated with this model to make sure if the user is reading a headline on a newspaper or a different language newspaper, he/she should be able to know the price or movement of a stock just by clicking the picture of the headline on a newspaper of any language and upload that picture on the portal.

Also the user can have a quick view on the real time stock history or stock prices jut by selecting the ticker and the no of days/months the user wants to see, the model will return real time graph of the selected day/month for stock price of the particular company of which the ticker has been selected. Also the system provides an option to upload the news headlines as well as the whole web-app in three different languages which are English, Hindi and Marathi.

The section I of the paper explains the introduction of general stock movement prediction using classification methods such as Random Forest Classifier. Section II presents the literature review of system and Section III presents proposed system architecture Section IV presents an evaluation parameter used for calculating the accuracy of the model. Section V provides us with the results. Section VI gives the conclusion, whereas at the end references and links are presented.

2. LITERATURE REVIEW

Stock Market Prediction using “Sentiment Analysis” has been researched for decades and is rising in its popularity. The paper [1] presented by John Kordonis proposed that he had developed a system which feeds past twitter tweets as input, process them using machine learning algorithms such as Naive

Bayes Classifier and Support Vector Machine and gives sentiment analysis on Stock Price Movements. This paper reflects the direct relation between public sentiment and stock market.

In the literature [2] by Arti Buche great efforts have been taken to predict stock market using financial news. It performed text opinion mining which gives sentiment analysis of stock price movements. The algorithms used were combinations of statistical based predictions and NLP based predictions which gave polarity of positive, negative and neutral sentiments.

Ayman E Khedr and S.E. Salama in their research [3] proposed a model by analyzing financial news articles and historical prices for predicting the stock market behaviour. It was conducted in two stages: 1st stage was to determine the sentiments of news articles using Naive Bayes algorithm. The output of 1st stage is then fed as input to 2nd stage along with historical attributes, both Naive Bayes and K-NN methods gave suitably accurate results.

In the cited paper [4] by Kalyani Joshi studies how news articles and stock prices can be related to each other. They developed a system for text mining using Dictionary based approach i.e. Bag Of Word Technique. The algorithms used were Random Forest Classifier, Support Vector Machine and Naïve Bayes Classifier, Out of which RF and SVM performed well with suitable accuracy.

The literature review [5] by Anshul Mittal proposed a system which correlates Public Sentiment and Market Sentiment. They performed Sentiment Analysis on Twitter Data to classify the public mood as Calm, Happy, Alert and Kind. These moods were then used to predict the future stock movements and further used the predicted values as a solution for the investors to have a clear idea whether to buy or sell a particular stock. It is inferred from the research that people’s mood indeed affected investment decision.

Ashish Pathak’s work [6] was based on News and Twitter Headlines to predict the sentiment of a company using text mining and Sentiment Analysis. In this, accurate results could be obtained to predict the market trend which proved that modern approach out performed Traditional approach.

The research paper [7] by Dev Shah, was able to successfully forecast the stock market trend by developing a Sentiment Analysis using Dictionary based approach for the pharmaceutical market sector and predict the news polarity as positive, negative or neutral and hence the Stock Market as up or down. The Dictionary

has 100 words for each news article. ”Pattern” was used which is a Python Library, for changing text corpus into numerical vectors i.e n grams. Thus it achieved a suitable accuracy of 70.59%.

The research document [8] by Rajesh K Ahir and Mittal B Ahir aims to give a comprehensive account of how Sentiment Analysis can be used to predict stock market movements. This paper presented a systematic approach based on text mining using algorithms like Naïve Bayes, kNN classifier and Decision Tree which gave appropriate results.

Pranjal Chakraborthy, in the research report [9] used Sentiment Analysis model and concluded that SVM worked best on the training data with unigram feature which was used to predict the future movement of US (Dow Jones)by analyzing Twitter Sentiment of Stock Market.

Anurag Nagar and Micheal Hahsler [10] aggregated news from various sources then thereby creating a News Corpus, which was further used and filtered to particular sentences and then NLP techniques were applied.

It was shown [11] by Yoosin Kim and Seung Ryul Jeong in their work that introduced a technique of mining text opinions in order to analyze Korean language News to predict rise and fall in KOSPI i.e Korea Composite Stock Price Index. They carried out the NPL of news, then describing its features and categorizing and extracting opinions and sentiments conveyed by the writers. Then a suitably accurate prediction about the relation between the stock market ups or downs and prediction was obtained.

3. SYSTEM ARCHITECTURE

A. Architecture Overview

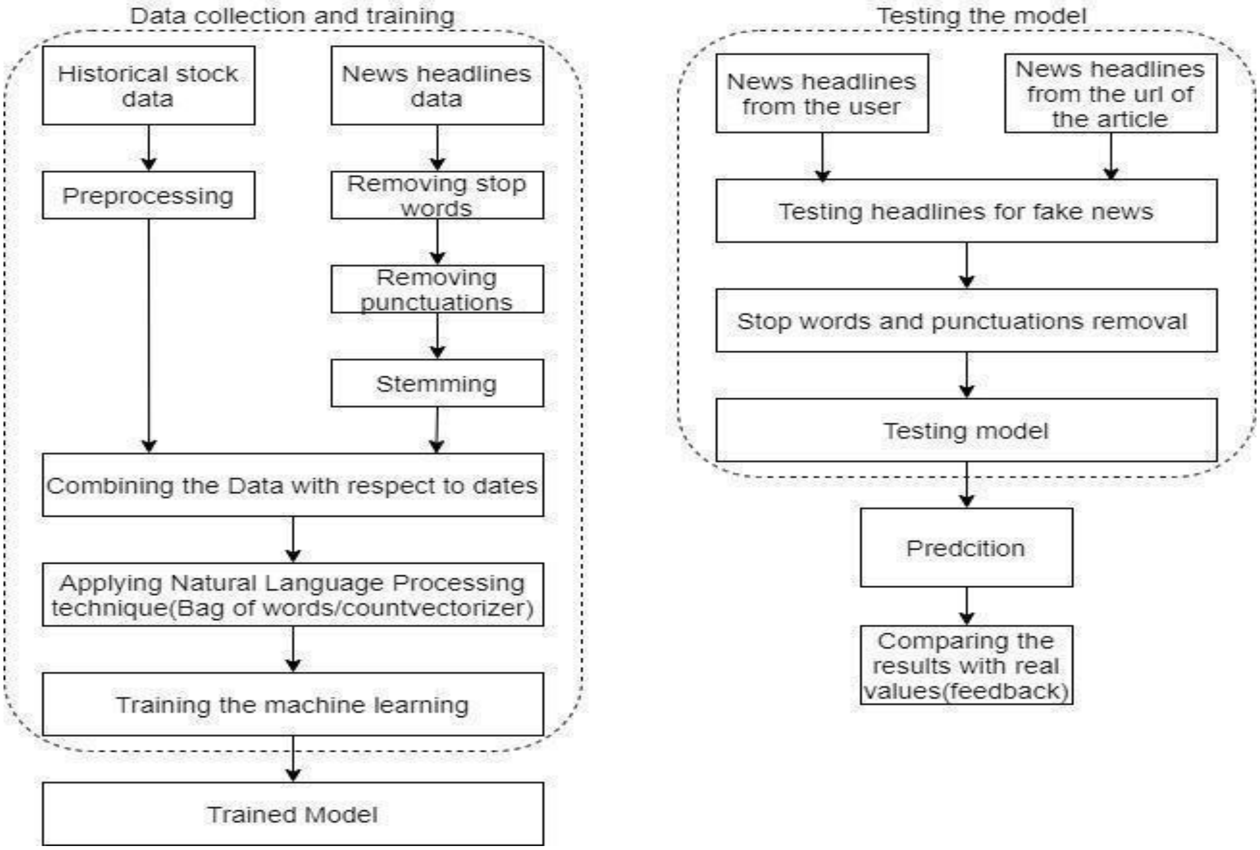

Data collection and training
Historical stock data
News headlines data
Preprocessing
Removing stop words
Removing punctuations
Stemming
Combining the Data with respect to dates
Applying Natural Language Processing technique(Bag of words/countvectorizer)
Training the machine learning
Trained Model
Testing the model
News headlines from the user
News headlines from the url of the article
Testing headlines for fake news
Stop words and punctuations removal
Testing model
Predcition
Comparing the results with real values(feedback)

Fig 1.a **Fig 1.b**

Fig 1. System Architecture

The fig 1.b shows the system architecture in which users will enter either the news headline itself or the URL of the news article, now the entered news headlines will initially be tested for being fake, if the entered headline is fake then the system will pop up a message saying 'The entered headline/news is a fake news. But if the entered headline/news is genuine then all the stop words as well as punctuations were removed and the very next moment it is been tested for our machine learning model which then predicted the results so far. The Fig 1.a shows the overview of the model for data collection and training the model in which the first step was collecting the historical stock data of several companies over the years and simultaneously the natural language processing methods were applied on the news headlines data in which we initially removes the stop words punctuations the next step was the stemming process which means reduction of inflected words to the base or root word for example – consider words such as ‘development’, ‘developed’, ‘developing’, all three mentioned words mean the same thing. So by using the process of stemming in natural language processing, all these three words will be considered as a same word and then the next step was combining data with respect to dates as in which headline came on which day. Then the next and the most important part of this process was applying natural language processing techniques such as bag of words, count vectorizer. Bag of words basically counts the no of words occurring in a sentence or a paragraph, for example - let us consider the input text as, ‘it was the best experience’, ‘it was the worst experience’, ‘it can work daily’, ‘it can't work daily’. The model will treat each line above as document and represent the occurrence of each word by making the list of all unique words. List of words--[it, was, the, best, experience, worst, can, work, daily, can't]. The model will represent each line as follows:' it was the best experience' = [1,1,1,1,1,0,0,0,0,0]'it was the worst experience'[1,1,1,0,1,1,0,0,0,0]' it can be used frequently' = [1,0,0,0,0,0,1,1,1,] 'it can't be used frequently' = [1,0,0,0,0,0,1,1,1]. The above representation of data is of type unigram (since the occurrence of only one word is considered). But the above representation of data is not helpful while classifying text like-'it was not good experience'. Because even though 'not good' is negative sentiment but it would be classified as positive due to the occurrence of word 'good'. So for such cases we do consider the vocabulary of grouped words. For example in bigram model we create the list of couple of word pairs such as, ['it was', 'was the', 'the best', 'best experience']. So, by using the bigram/trigram model, we can successfully detect the occurrence of negative sentiments.

B. Algorithm Used

The algorithms we used were Random Forest classifier, decision tree classifier, logistic regression, support vector machine (SVM).

Random Forest Classifier

Random Forest Classifier algorithm is an algorithm that creates set of decision tress and sums up the votes from different decision tress and then it finally decides the class of the test object. Let’s say the training set is given as [T1, T2, T3, T4] now the random forest will take input of subset and create three different decision tress as [X1, X2, X3], [X2, X3, X4], [X1, X2, X4] and now it will predict considering the majority of the votes each decision tree will make. A single decision tree may prone to a noise but aggregate of more decision trees will not, it will rather reduce the effect of noise resulting in more accurate results. Also random forests works well for a vast range of data items than a one decision tree does and the scaling of data does not require in random forest classifier as well. Feature Randomness — In a normal decision tree, when it is time to split a node, we basically consider each and every possible feature and then picking out the one that produces the most separation between the observations that were originally left node vs. the ones in the right node. In contrast, each and every tree in a random forest can only pick from a random feature subset. This actually forces and more diversification more variation amongst the trees in the model and finally results in lower correlation across trees. So in the random forest, we are left with the trees which are not only trained on various different sets of data but also which use different features which helps them to make decisions.

Day	Random Forest Classifier on 80%-20% split
Day 1	83.86%
Day 2	83.33%
Day 3	84.13%
Day 4	83.86%
Day 5	84.39%

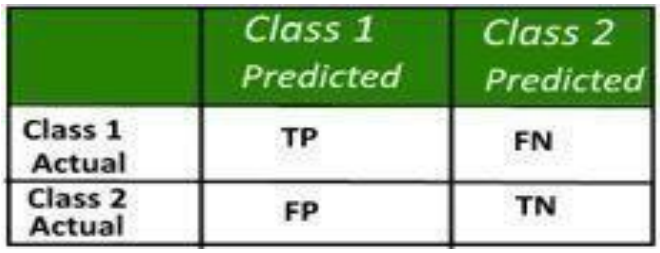
Day 6	83.6%
Day 7	83.33%
Day 8	83.86%
Day 9	84.13%
Day 10	83.07%
Average Accuracy over 10 Days	83.75%

Table 1. Average Classification Accuracy over the period of 10 days

The accuracy of the model for 10 different days is shown in the Fig 3. As shown in the figure, the average accuracy over the period of 10 days for the Random Forest Classifier on 80%20% split was 83.75%.

4. EVALUATION PARAMETERS

Evaluating the machine learning model is essential to test the capabilities of the algorithm used and the train, test split of available data set. This can be done by calculating the accuracy of the machine learning model the confusion matrix is used to calculate the accuracy. It is basically the ration of number of correct predictions upon total number of predictions made Accuracy = Confusion matrix is a 2x2 matrix which is the summary of prediction results on classification algorithm.



	Class 1 Predicted	Class 2 Predicted
Class 1 Actual	TP	FN
Class 2 Actual	FP	TN

Here class1: positive and class2: negative Definition of terms: True Positive that is (TP) states that an observation is positive, and also is supposed to be predicted positive; False Negative that is (FN) states that an observation is positive, but it predicts to be negative; True Negative that is (TN) states that an observation is negative, and is predicted to be negative; False Positive that is (FP) states that an observation is negative, whereas it predicts to be positive.

For example if we assume values TP=153, FN=33, FP=30, For example we assume values TP=153, FN=33, FP=30, TN=162 which is $Accuracy = {}^{TP+TN}/\ TP+FN+FP+TN$

$$Accuracy = \frac{153+162}{153+33+30+162} = 0.8334$$

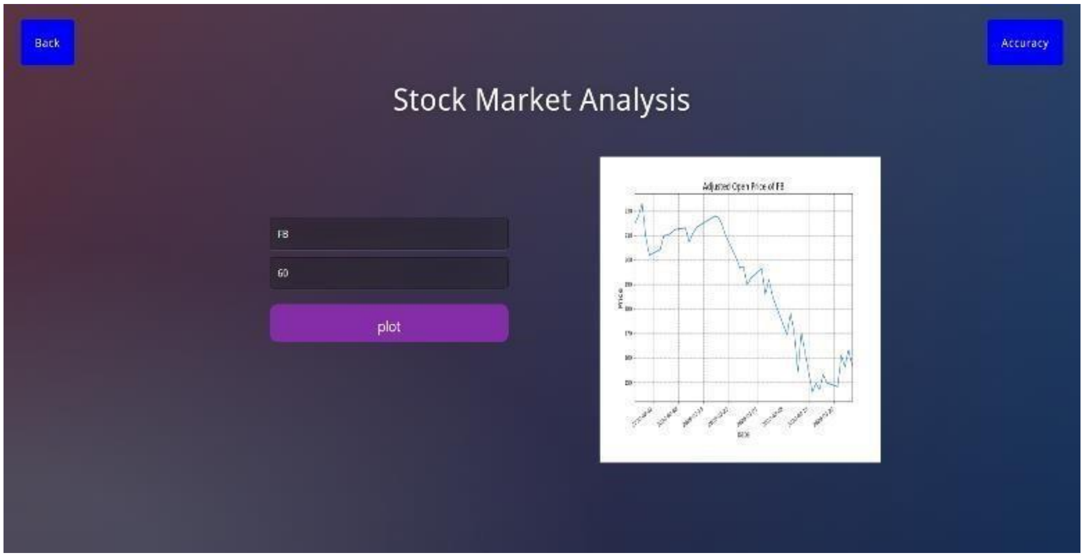
Which means the accuracy is 83.34%.

Table 2. Accuracy of different train set and test set split

Model	80% Data split	70% Data split
Random Forest Classifier	83.96%	58.52%
Decision Tree Classifier	81.21%	57.87%
Logistic Regression	82.01%	59.39%
Support Vector Machine	83.86%	61.41%

From all four used algorithms of which we tested all four algorithms on 80%-20% train set-test set split and on 70%-30% train set-test set split, in which we got the highest accuracy of random forest classifier with 83.96% accuracy followed by the support vector machine with 83.86% followed by logistic regression with 82.01% followed by decision tree classifier with 81.21% accuracy. All the accuracy results are shown in Fig 4.

5. RESULT



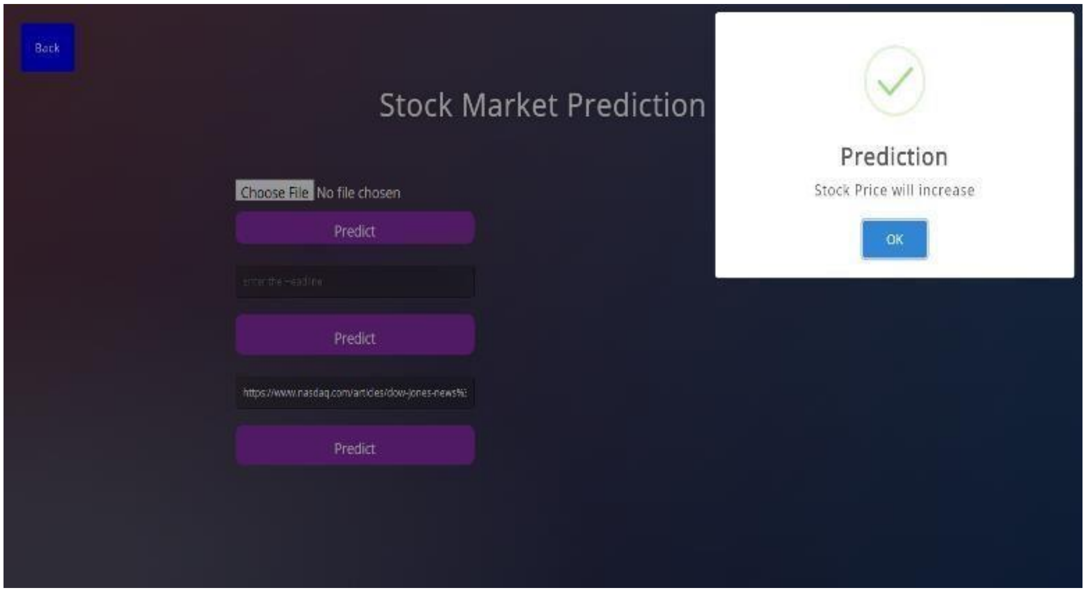
Stock Market Analysis

FB

60

Fig 2 Real time analysis of stock of Ticker 'FB'

The above figure shows the real time graph of the Facebook of over last 60 days, whenever the movement on the graph is rising it states the stock price has increased and when the movement of the graph is declining it states that the stock price of Facebook has decreased.



No file chosen

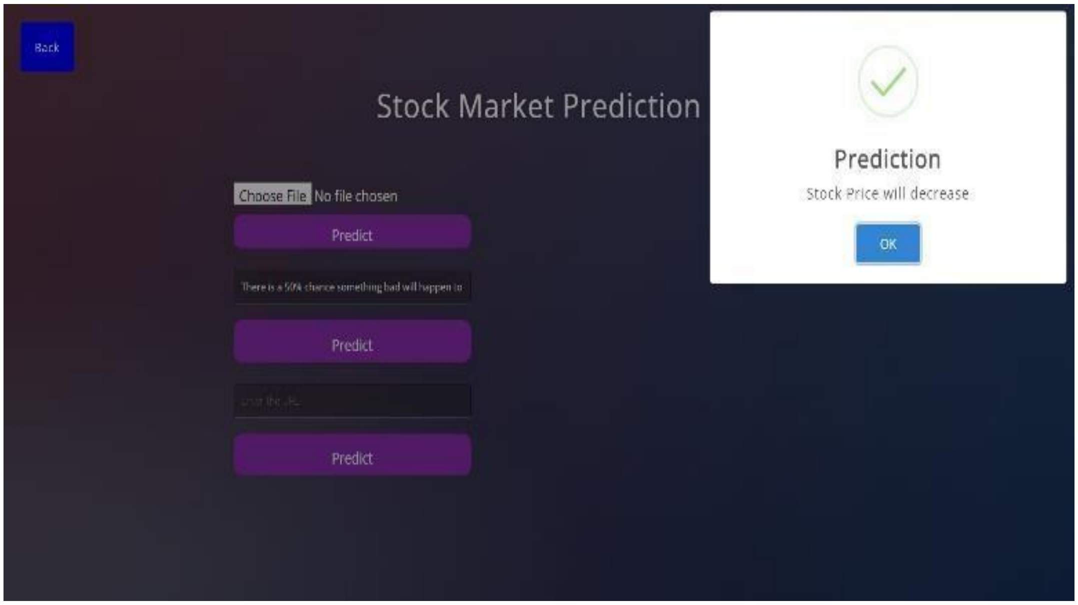
https://www.nasdaq.com/articles/dow-jones-news%:

Prediction

Stock Price will increase

Fig 3 Predicting stock movement on entering a positive article URL

The above figure shows that on entering the article URL link as 'https://www.nasdaq.com/articles/dow-jones-news%3Aappleproduction-problemsboeing-upgraded-2020-02-10l', it fetches the article from the provided URL and then summarizes it, on the basis of which the model predicts the movement of stock.



No file chosen

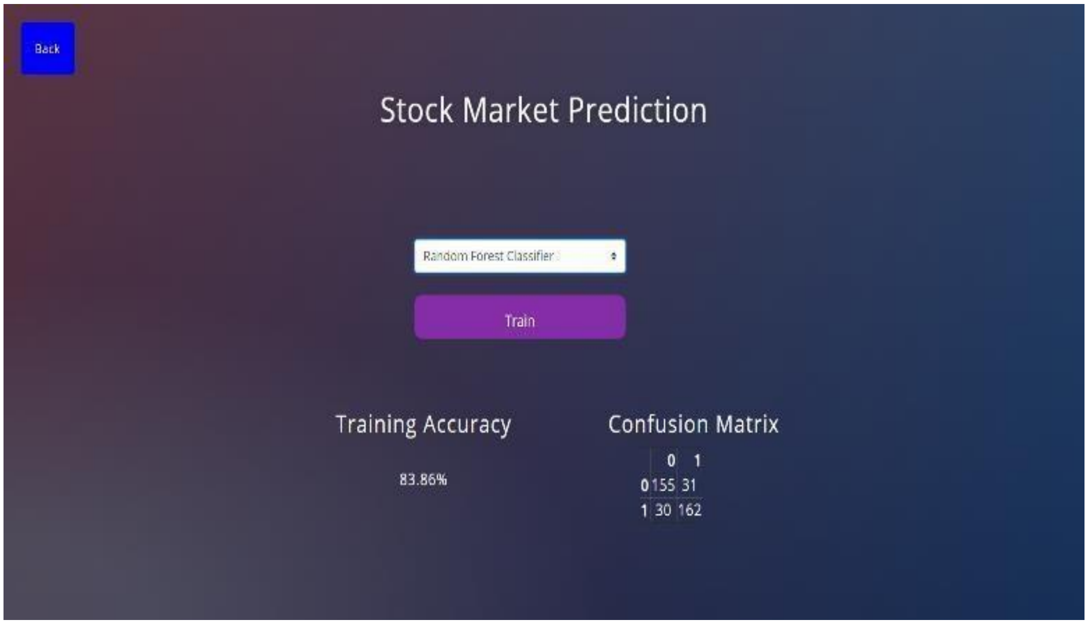
There is a 50% chance something bad will happen to

Prediction

Stock Price will decrease

Fig 4 Prediction of Stock on entering a negative headline

The above figure shows that on entering the news headline as 'There was a 50% something bad will happen around the southern region ' , on the basis of which the model predicts the movement of stock.

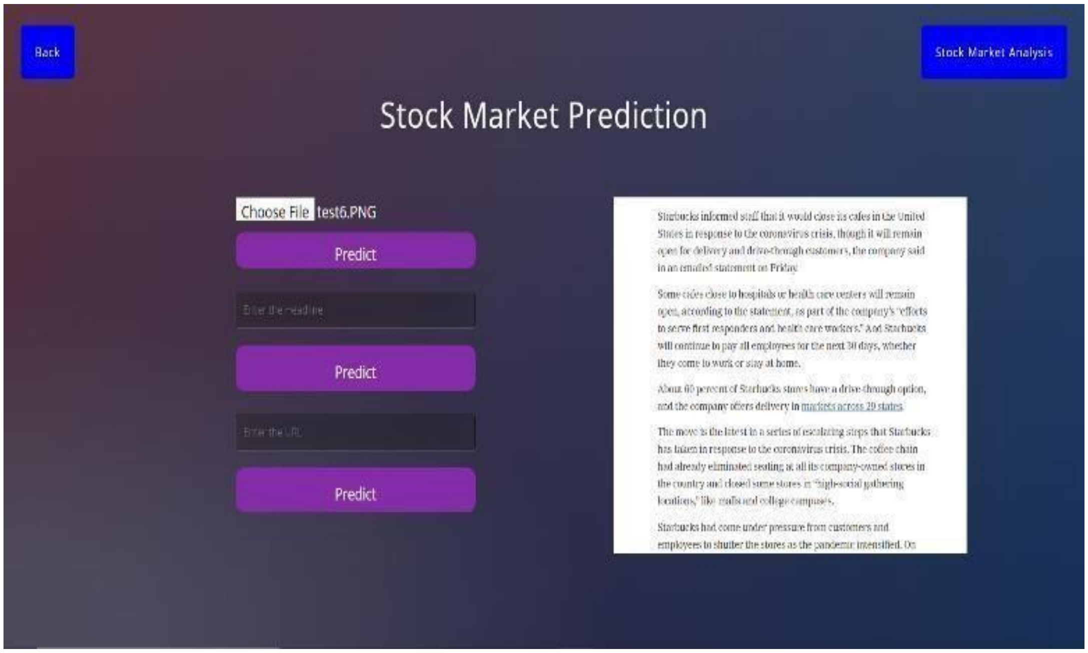


83.86%

	0	1
0	155	31
1	30	162

Fig 5 Training Accuracy and Confusion matrix of selected algorithm which is Random Forest Classifier

The above figure shows the Training Accuracy and the Confusion Matrix of the selected algorithm, on choosing Random Forest Classifier, we get 83.33% accuracy.



test6.PNG

Starbucks informed staff that it would close its cafes in the United States in response to the coronavirus crisis, though it will remain open for delivery and drive-through customers, the company said in an emailed statement on Friday.

Some cafes close to hospitals or health care centers will remain open, according to the statement, as part of the company's "efforts to serve first responders and health care workers." And Starbucks will continue to pay all employees for the next 30 days, whether they come to work or stay at home.

About 60 percent of Starbucks stores have a drive-through option, and the company offers delivery in markets across 20 states.

The move is the latest in a series of escalating steps that Starbucks has taken in response to the coronavirus crisis. The coffee chain had already eliminated seating at all its company-owned stores in the country and closed some stores in "high-social gathering locations," like malls and college campuses.

Starbucks had come under pressure from customers and employees to shutter the stores as the pandemic intensified. On

Fig 6 Uploading the image file which gets converted to text with the help of OCR

6. CONCLUSION

We proposed a stock movement prediction model based on machine learning algorithms with the web-app in different languages such as English, Hindi and Marathi. We utilized Random Forest Classifier algorithm to classify the news headlines and news articles data because stock trends are very volatile and it is very complicated to predict the future stock trends and compare it with real time analysis of any particular company.

REFERENCES

[1] Kordonis, J. , &Symeonidis, S. Stock Market Price Forecasting via sentimental analysis on Twitter. Working paper, The 20th Panhellic Conference on Informatics (PCI'16), 10-12 November.
[2] Arti Buche and Dr M.B. Chandak,Stock Market Forecasting Techniques: A Survey, Journal of Engineering and Applied Sciences 14 (5): 1649-1655, 2019 ISSN: 1816-949X
[3] A. E. Khedr and N. Yaseen, Prediction of Stock Market Behaviour via Data Mining Technique and News Sentimental Analysis, International Journal of Intelligent Systems and Applications 9(7):22-30 DOI: 10.5815/ijisa.2017.07.03
[4] Joshi, K. , &Prof. Bharathi H. N., Stock Trend Prediction Using News Sentimental Analysis, International Journal of Computer Science and Information Technology (IJCSIT) Volume 8 , No 3, June (2016).
[5] Mittal, A. , &Goel, A(2011), Stock Prediction by Twitter Sentimental Analysis, Standford University,CS229.
[6] Pathak, A. , &Shetty N.P(2019)Indian Stock Market Forecast by ML and Sentimental Analysis, Computational Intelligence in Data Mining pp595-603.
[7] D.Shah and H.Isah,"Effects of News Sentiment for prediction of Stock Market"2018 IEEE International Conference on Big Data(Big Data)Seattle.WA,USA,2018,pp.4705-4708.
[8] Rajesh K Ahir and Mittal B Ahir,"Predictive Sentimental Analysis of Stock Tweets". international journal of modern trends in engineering and technology vol.(8) issue(1),pp.071-077
[9] Pranjal Chakraborthy ,"Sentimental Analysis of Twitter Feed",2017.6th International conference on Informatics, Electronics and Vision.
[10] Nagar, A. , &Hahsler, M.(2012) ,"Using text mining techinque and data mining methods to extract stock market sentiments from live news stream",2012. international conference on computer technology and science (ICCTS2012)IPCSITvol.XX(2012).
[11] Kim, Y. , &Jeong, S. R. (2014),"Text opinion mining to analyse news for Stock Market Forecast',Int.J.Advanced.Soft Computing Applied,Volume 6.No.1.March 2014,ISSN 2074/8523.