International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 5 Volume No.7 September - October – 2023
DOI:10.46647/ijetms.2023.v07i05.003 ISSN: 2581-4621

Stock Price Prediction Using Twitter Dataset

M.Reddy Vijay Kumar[1], L.Gopi Krishna[2]
[1]*M.Tech Student, Department of CSE, Golden Valley Integrated Campus, Madanapalli*
[2]*Associate Professor, Department of CSE, Golden Valley Integrated Campus, Madanapalli*

Abstract:
In this paper, we apply sentiment analysis and machine learning principles to find the correlation between ”public sentiment” and ”market sentiment”. We use twitter data to predict public mood and use the predicted mood and previous days’ DJIA values to predict the stock market movements. In order to test our results, we propose a new cross validation method for financial data and obtain 75.56% accuracy using Self Organizing Fuzzy Neural Networks (SOFNN) on the Twitter feeds and DJIA values from the period June 2009 to December 2009. We also implement a naive protfolio management strategy based on our predicted values. Our work is based on Bollen et al’s famous paper which predicted the same with 87% accuracy.
Keywords: Stock, Prediction, analysis, twitter

INTRODUCTION

Stock market prediction has been an active area of research for a long time. The Efficient Market Hypothesis (EMH) states that stock market prices are largely driven by new information and follow a random walk pattern. Though this hypothesis is widely accepted by the research community as a central paradigm governing the markets in general, several people have attempted to extract patterns in the way stock markets behave and respond to external stimuli.

In this paper, we test a hypothesis based on the premise of behavioral economics, that the emotions and moods of individuals affect their decison making process, thus, leading to a direct correlation between ”public sentiment” and ”market sentiment”. We perform sentiment analysis on publicly available Twitter data to find the public mood and the degree of membership into 4 classes - Calm, Happy, Alert and Kind (somewhat like fuzzy membership). We use these moods and previous days’ Dow Jones Industrial Average (DJIA) values to predict future stock movements and then use the predicted values in our portfolio management strategy

Related work Our work is based on Bollen et al’s strategy [1] which received widespread media coverage recently. They also attempted to predict the behavior of the stock market by measuring the mood of people on Twitter. The authors considered the tweet data of all twitter users in 2008 and used the OpinionFinder and Google Profile of Mood States (GPOMS) algorithm to classify public sentiment into 6 categories, namely, Calm, Alert, Sure, Vital, Kind and Happy. They cross validated the resulting mood time series by comparing its ability to detect the public’s response to the presidential elections and Thanksgiving day in 2008. They also used causality analysis to investigate the hypothesis that public mood states, as measured by the OpinionFinder and GPOMS mood time series, are predictive of changes in DJIA closing values. The authors used Self Organizing Fuzzy Neural Networks to predict DJIA values using previous values. Their results show a remarkable accuracy of nearly 87% in predicting the up and down changes in the closing values of Dow Jones Industrial Index (DJIA).

The rest of the paper is organized as follows. The second section briefly discusses our general approach towards solving the problem and the following sections discuss the individual components in greater detail. In Section 3, we briefly discuss the dataset that we have used for this paper and data preprocessing measures adopted. Section 4 discusses the sentiment analysis technique developed by us for the purpose of this paper. Section 5 includes in detail, the different machine learning techniques to predict DJIA values using our sentiment analysis results

DOI:10.46647/ijetms.2023.v07i05.003 ISSN: 2581-4621

and presents our findings. In Section 6, we use the predicted values and devise a naive strategy to maintain a profitable portfolio.

LITERATURE SURVEY

Twitter mood as a stock market predictor :

Behavioral economics tells us that emotions can profoundly affect individual behavior and decision-making. Does this also apply to societies at large, i.e. can societies experience mood states that affect their collective decision making? By extension is the public mood correlated or even predictive of economic indicators? Here we investigate whether measurements of collective mood states derived from large-scale Twitter feeds are correlated to the value of the Dow Jones Industrial Average (DJIA) over time. We analyze the text content of daily Twitter feeds by two mood tracking tools, namely OpinionFinder that measures positive vs. negative mood and Google-Profile of Mood States (GPOMS) that measures mood in terms of 6 dimensions (Calm, Alert, Sure, Vital, Kind, and Happy). We cross-validate the resulting mood time series by comparing their ability to detect the public's response to the presidential election and Thanksgiving day in 2008. A Granger causality analysis and a Self-Organizing Fuzzy Neural Network are then used to investigate the hypothesis that public mood states, as measured by the OpinionFinder and GPOMS mood time series, are predictive of changes in DJIA closing values. Our results indicate that the accuracy of DJIA predictions can be significantly improved by the inclusion of specific public mood dimensions but not others. We find an accuracy of 87.6% in predicting the daily up and down changes in the closing values of the DJIA and a reduction of the Mean Average Percentage Error by more than 6%.

A library for support vector machines :

LIBSVM is a library for Support Vector Machines (SVMs). We have been actively developing this package since the year 2000. The goal is to help users to easily apply SVM to their applications. LIBSVM has gained wide popularity in machine learning and many other areas. In this article, we present all implementation details of LIBSVM. Issues such as solving SVM optimization problems theoretical convergence multiclass classification probability estimates and parameter selection are discussed in detail.

An on-line algorithm for creating self-organizing fuzzy neural networks :

This paper presents a new on-line algorithm for creating a self-organizing fuzzy neural network (SOFNN) from sample patterns to implement a singleton or Takagi-Sugeno (TS) type fuzzy model. The SOFNN is based on ellipsoidal basis function (EBF) neurons consisting of a center vector and a width vector. New methods of the structure learning and the parameter learning, based on new adding and pruning techniques and a recursive on-line learning algorithm, are proposed and developed. A proof of the convergence of both the estimation error and the linear network parameters is also given in the paper. The proposed methods are very simple and effective and generate a fuzzy neural model with a high accuracy and compact structure. Simulation work shows that the SOFNN has the capability of self-organization to determine the structure and parameters of the network automatically.

ALGORITHM

The technique used in this paper builds directly on the one used by Bollen et al. [1]. The raw DJIA values are first fed into the preprocessor to obtain the processed values. At the same time, the tweets are fed to the sentiment analysis algorithm which outputs mood values for the four mood classes for each day. These moods and the processed DJIA values are then fed to our model learning framework which uses SOFNN to learn a model to predict future DJIA values using them. The learnt model as well as the previous DJIA and mood values are used by the portfolio management system which runs the model to predict the future value and uses the predicted values to make appropriate buy/sell decisions. Figure 1 shows a brief flow diagram of our technique.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 5 Volume No.7 September - October – 2023
DOI:10.46647/ijetms.2023.v07i05.003 ISSN: 2581-4621

DATASET

In this project, we used two main datasets-

• Dow Jones Industrial Average (DJIA) values from June 2009 to December 2009. The data was obtained using Yahoo! Finance and includes the open, close, high and low values for a given day.

• Publicly available Twitter data containing more than 476 million tweets corresponding to more than 17 million users from June 2009 to December 2009. The data includes the timestamp, username and tweet text for every tweet during that period. Since we perform our prediction and analysis on a daily basis, we split the tweets by days using the timestamp information.

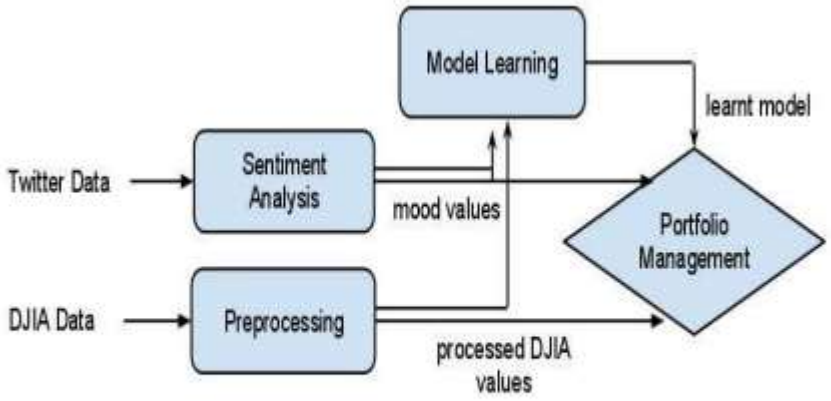


Figure 1: Our technique

Data Preprocessing The data obtained from the above mentioned sources had to be pre-processed to make it suitable for reliable analysis. We pre-processed the DJIA data in the following manner.

1. While the Twitter data was available for all days lying in the giving period, the DJIA values obtained using Yahoo! Finance was (understandably) absent for weekends and other holidays when the market is closed. In order to complete this data, we approximated the missing values using a concave function. So, if the DJIA value on a given day is x and the next available data point is y with n days missing in between, we approximate the missing data by estimating the first day after x to be (y+x)/2 and then following the same method recursively till all gaps are filled. This approximation is justified as the stock data usually follows a concave function, unless ofcourse at anomaly points of sudden rise and fall.
2. If we observe the general movement of stock markets, it is associated with a few sudden jumps/falls and a brief period of small fluctuations around the new value. However, such jumps/falls are due to some major aberrations and cannot be predicted. Moreover, as we know the public memory is very short and even though the market may be trading at a much higher level than the previous year, that does not mean that calmness will be much higher than previous year; public mood is a very local metric. Therefore, we adjusted our stock values by shifting up/down for steep falls/jumps, respectively; making sure that we do not disturb the daily directional trend (up/down movement of stock prices).
3. Even after shifting the values in step 2, the values contained significant periods of volatile activity which are very difficult to predict. We pruned our dataset by removing these periods for final training and testing.

Finally, in order to ensure that values were small and comparable, we computed the z-score of each point in the data series ($(x-\mu)/\sigma$) and used that in our analysis (The original values were of the order of 104 , so MATLAB was giving a precision error when computing functions like $\exp(-x\ 2\ ))$

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 5 Volume No.7 September - October – 2023
DOI:10.46647/ijetms.2023.v07i05.003 ISSN: 2581-4621

SENTIMENT ANALYSIS

Sentiment analysis was an important part of our solution since the output of this module was used for learning our predictive model. While there has been a lot of research going on in classifying a piece of text as either positive or negative, there has been little work on multi-class classification. In this project, we use four mood classes, namely, Calm, Happy, Alert, and Kind. We tried several standard tools like OpinionFinder, SentiWordnet [5] etc. for our problem but found them inadequate and/or inefficient and therefore decided to develop our own analysis code. The methodology we adopted in finding the public sentiment is as follows

1. Word List Generation :

We develop our own word list based on the well known Profile of Mood States (POMS) questionnaire. POMS is an established psychometric questionnaire which asks a person to rate his/her current mood by answering 65 different questions on a scale of 1 to 5 (For example, rate on a scale of 1 to 5 how tensed you feel today?). These 65 words are then mapped on to 6 standard POMS moods- Tension, Depression, Anger, Vigour, Fatigue and Confusion. In order to do automate this analysis for tweets, the word list needs to be appropriately extended. Bollen et al. [1] used the Google n-grams data for the same. We followed a much simpler approach of extending the list by considering all commonly occuring synonyms of the base 65 words using SentiWordNet and a standard Thesaurus.

2. Tweet Filtering :

As mentioned earlier, the tweet data is enormous and will take several hours to be processed if used as it is (which makes the task of daily predictions difficult). Therefore, we filtered and considered only those tweets which are more likely to express a feeling, i.e. we consider only those tweets which contain the words ”feel”, ”makes me”, ”I’m” or ”I am” in them.

3. Daily Score Computation

We used a simple word counting algorithm to find the score for every POMS word for a given day

$$\text{score of a word} = \frac{\#\text{of times the word matches tweets in a day}}{\#\text{of total matches of all words}}$$

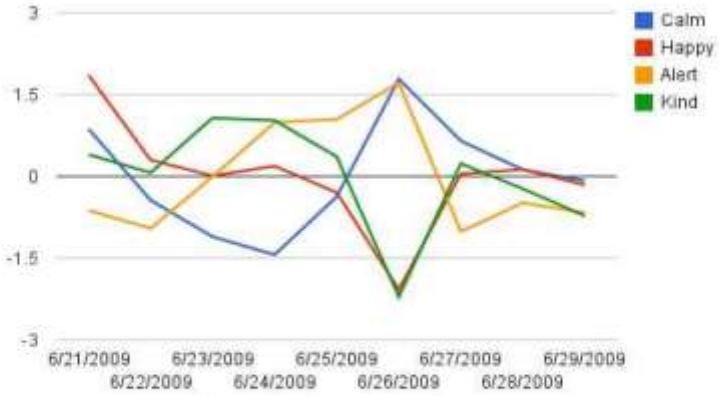
The denominator accounts for the fact that the number of tweets could vary from one day to another. This works well for our problem because of the nature of tweets which contain simple sentence structures and only a maximum of 140 characters (in most cases much less). We tried using the Stanford coreNLP software for word tagging and then using a word’s position in the sentence to find its importance. However, similar to our experience working with OpinionFinder, we observed that this process, besides being extremely slow was not too beneficial.

4. Score Mapping:

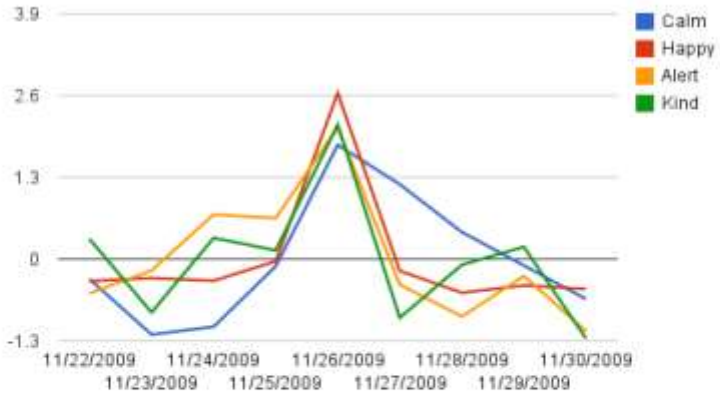
We map the score of each word to the six standard POMS states using the mapping techniques specified in the POMS questionnaire. We then map the POMS states to our four mood states using static correlation rules (for example, happy is taken as sum of vigour and negation of depression).

It is important to note that, given our formulation, it does not make much sense to compare the value of one mood against another; they should only be used to compare mood trends across days. We cross validated the results of our sentiment analysis technique by comparing the values returned by our algorithm around significant events like Thanksgiving day and Michael Jackson’s death. As shown in Figure 2, the moods show a sharp rise of various mood states on Thanksgiving whereas one day after MJ’s death, there is a sharp decline in happiness.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 5 Volume No.7 September - October – 2023
DOI:10.46647/ijetms.2023.v07i05.003 ISSN: 2581-4621



(a) Various moods after Michael Jackson's death on 25 June 2009



(b) Various moods on Thanksgiving day on 26 November 2009

Figure 2: Cross validation of our sentiment analysis by analyzing moods on some important events

Granger Causality In order to ascertain whether the mood values returned by our algorithm can be used to predict the future stock movements, we computed the p-values using Granger Causality analysis. Granger Causality analysis finds how much predictive information one signal has about another over a given lag period. The p-value measures the statistical significance of our result i.e. how likely we could obtain the causality value by random chance; therefore, lower the p-value, higher the predictive ability. It is clear from Table 1 that calmness and happiness are most helpful in predicting the DJIA values as per the Granger causality. Moreover, the causality is best when we use the past 3 or 4 days' data. Figure 3 shows the DJIA values superimposed with sum of calmness and happiness over a brief period, indicating the clear predictive trend.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 5 Volume No.7 September - October – 2023
DOI:10.46647/ijetms.2023.v07i05.003 ISSN: 2581-4621

Table 1: p-values obtained using Granger causality analysis with different lags (in days)

Lag	Calm	Happy	Alert	Kind
1	0.0207	0.4501	0.0345	0.0775
2	0.0336	0.1849	0.1063	0.1038
3	0.0106	0.0658	0.1679	0.1123
4	0.0069	0.0682	0.3257	0.1810
5	0.0100	0.0798	0.1151	0.1157

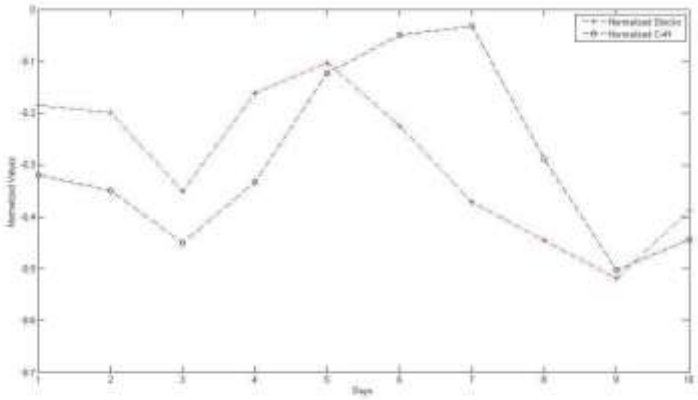
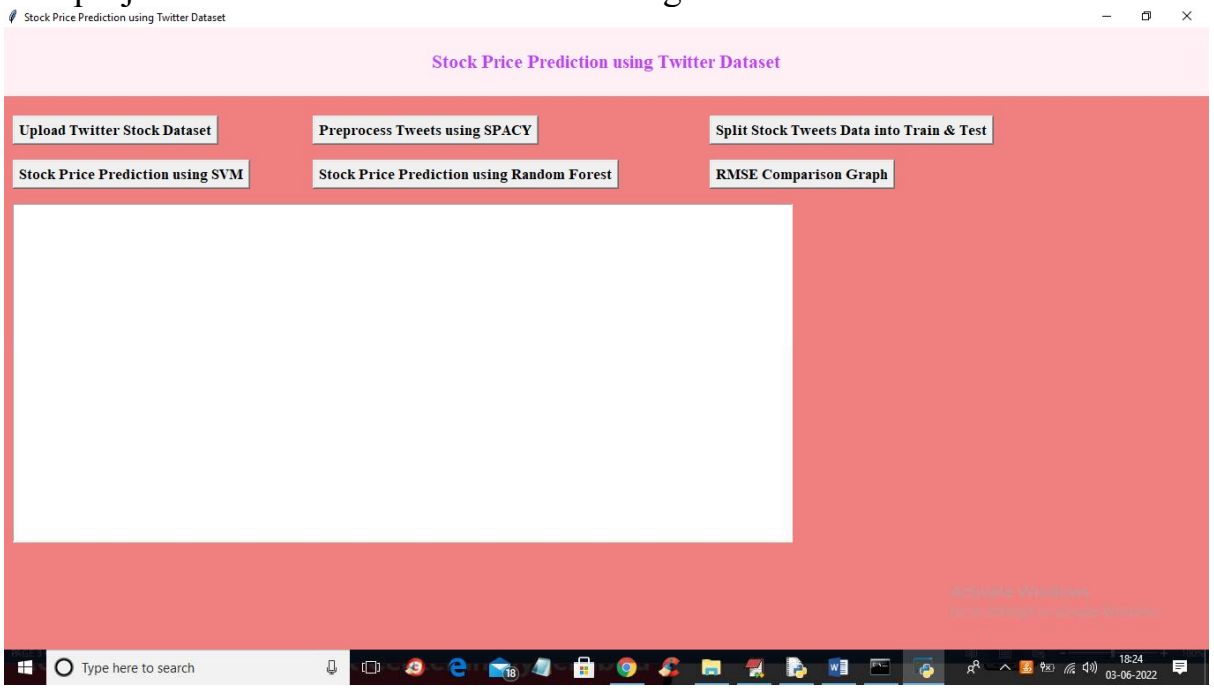


Figure 3: The DJIA, and Calmness + Happiness curves superimposed to show correlation

In the next section, we use the results of our sentiment analysis algorithm to learn a model that can predict the stock index and its movement.

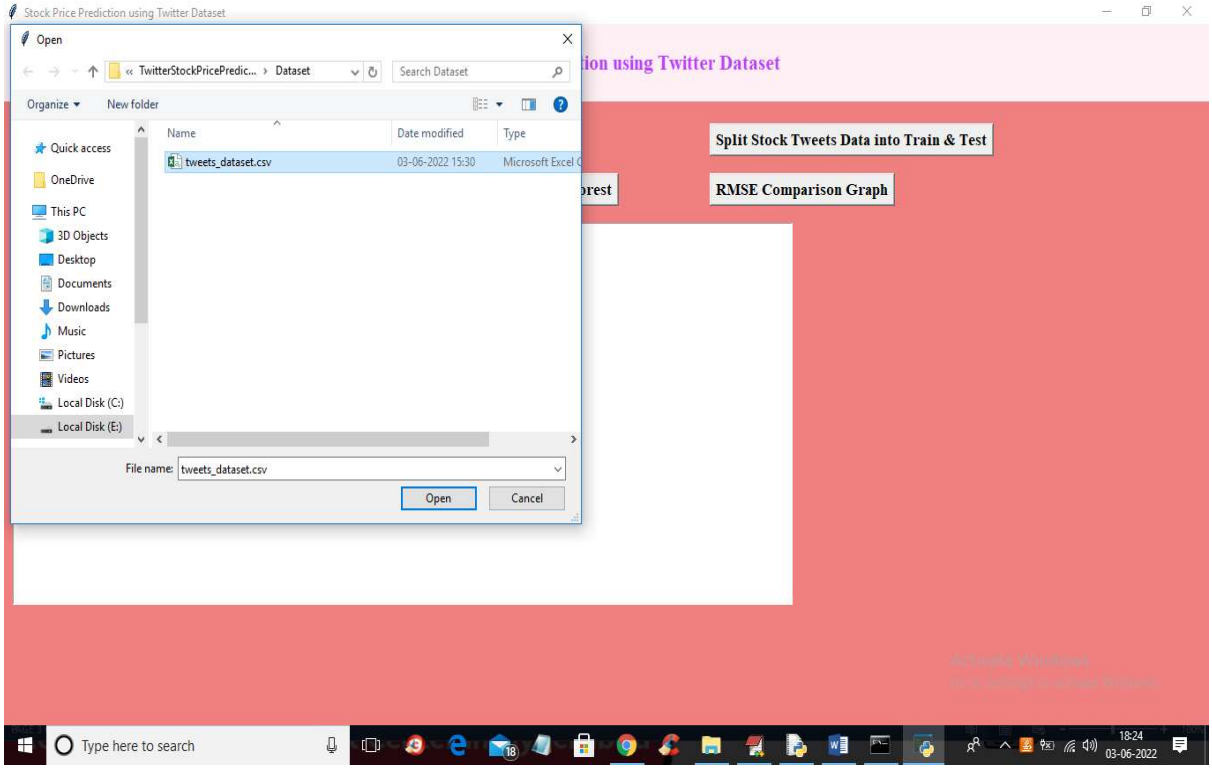
RESULTS

To run project double click on 'run.bat' file to get below screen

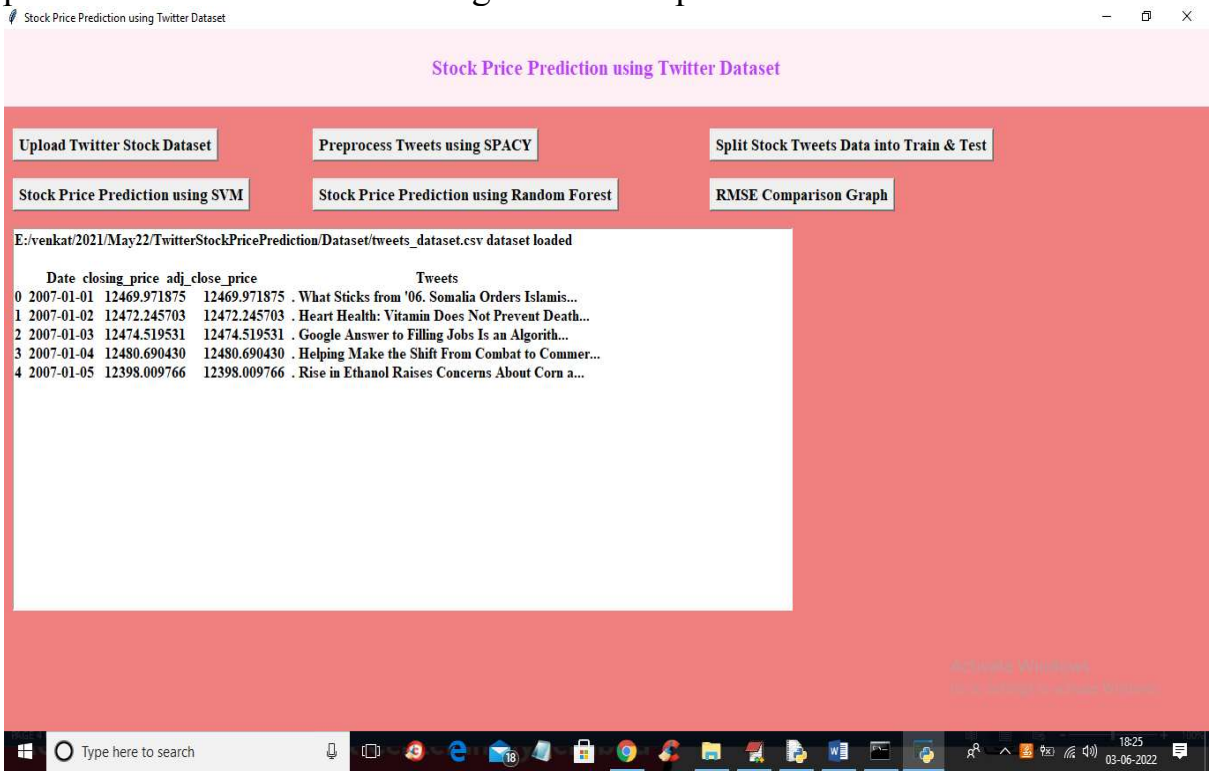


International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 5 Volume No.7 September - October – 2023
DOI:10.46647/ijetms.2023.v07i05.003 ISSN: 2581-4621

In above screen click on 'Upload Twitter Stock Dataset' button to upload dataset and get below screen

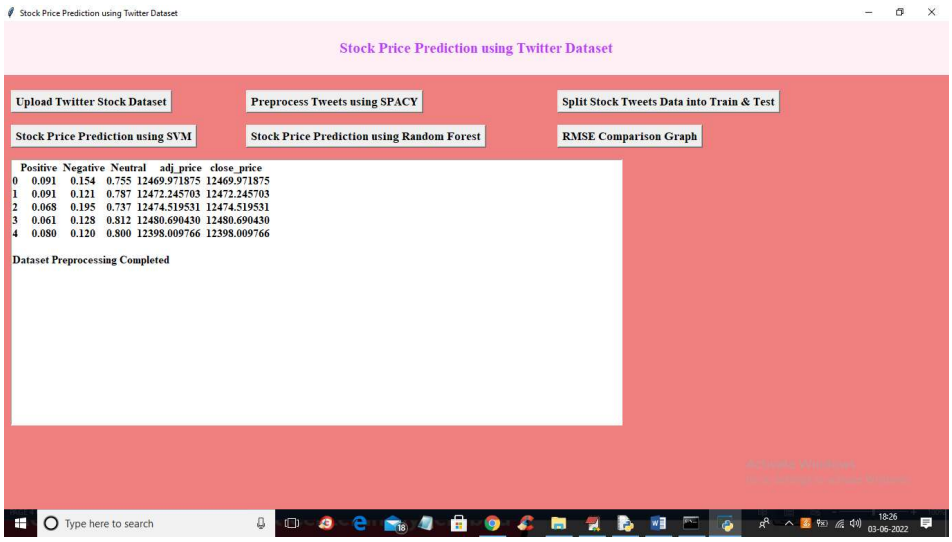


In above screen selecting and uploading tweets stock price dataset and then click on 'Open' button to load dataset and get below output

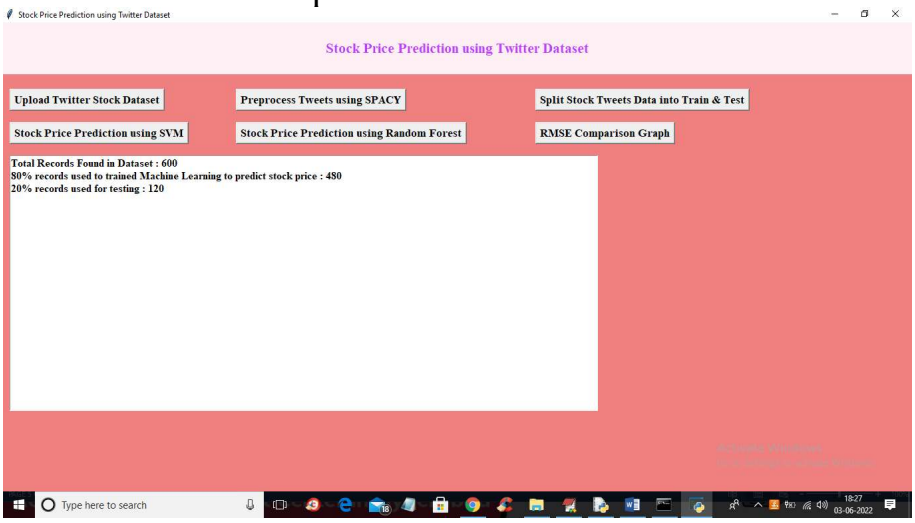


In above screen tweets and stock prices loaded and now click on 'Preprocess Tweets using SPACY' button to read all tweets and then find sentiment probability of each tweet in terms of positive, negative and neutral and then will get below output

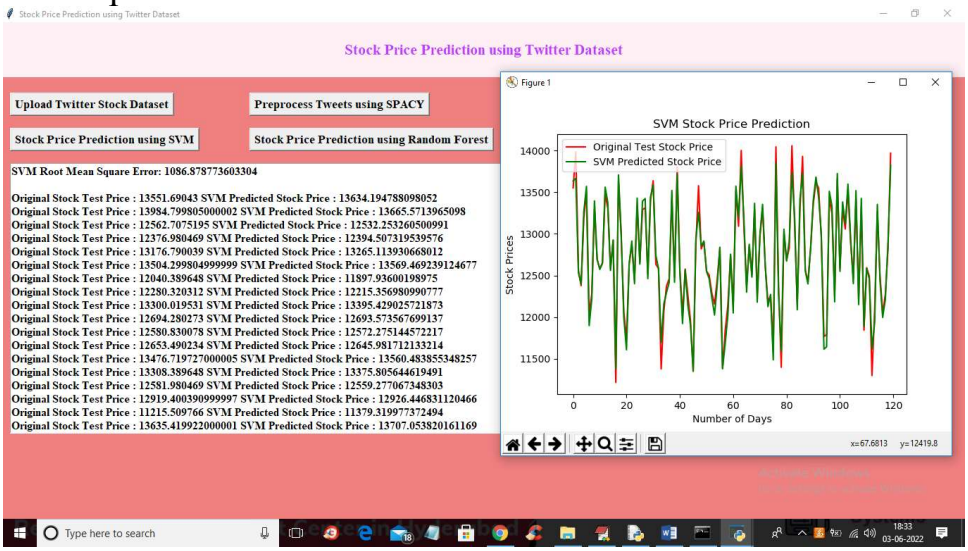
International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 5 Volume No.7 September - October – 2023
DOI:10.46647/ijetms.2023.v07i05.003 ISSN: 2581-4621



In above screen for tweet we got customer experience in terms of sentiments of stock performance and now click on 'Split Stock Tweets Data into Train & Test'

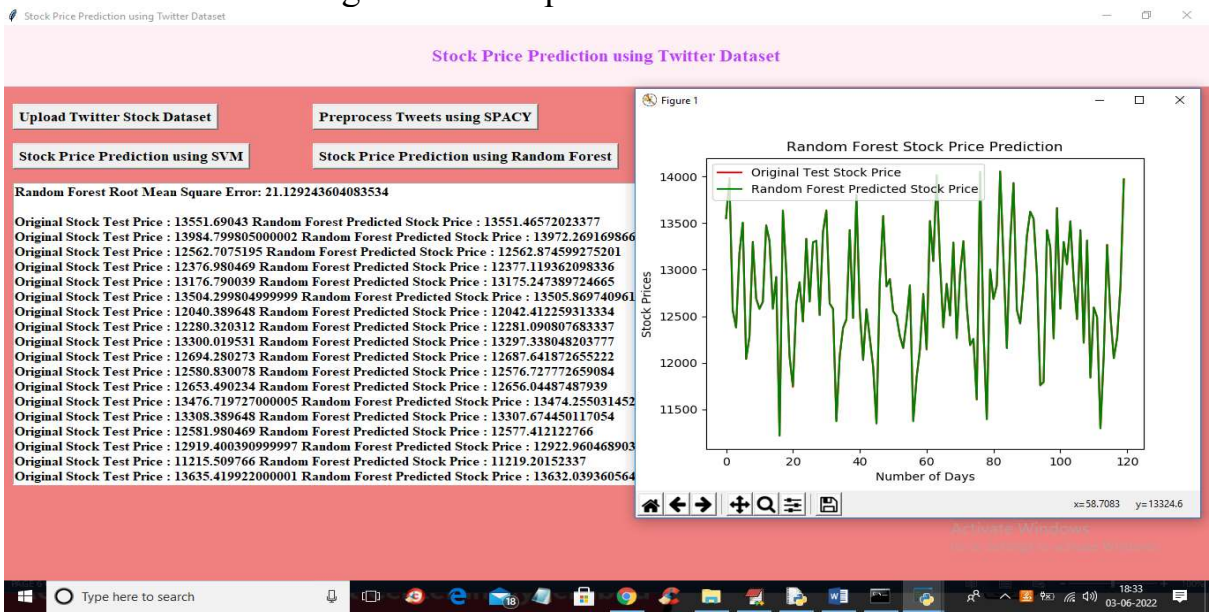


In above screen we can see dataset contain 600 tweets and application using 80% (480) tweets for training and 20% (120) tweets for testing and now train and test data is ready and now click on 'Stock Price Prediction using SVM' button to train SVM and get below output



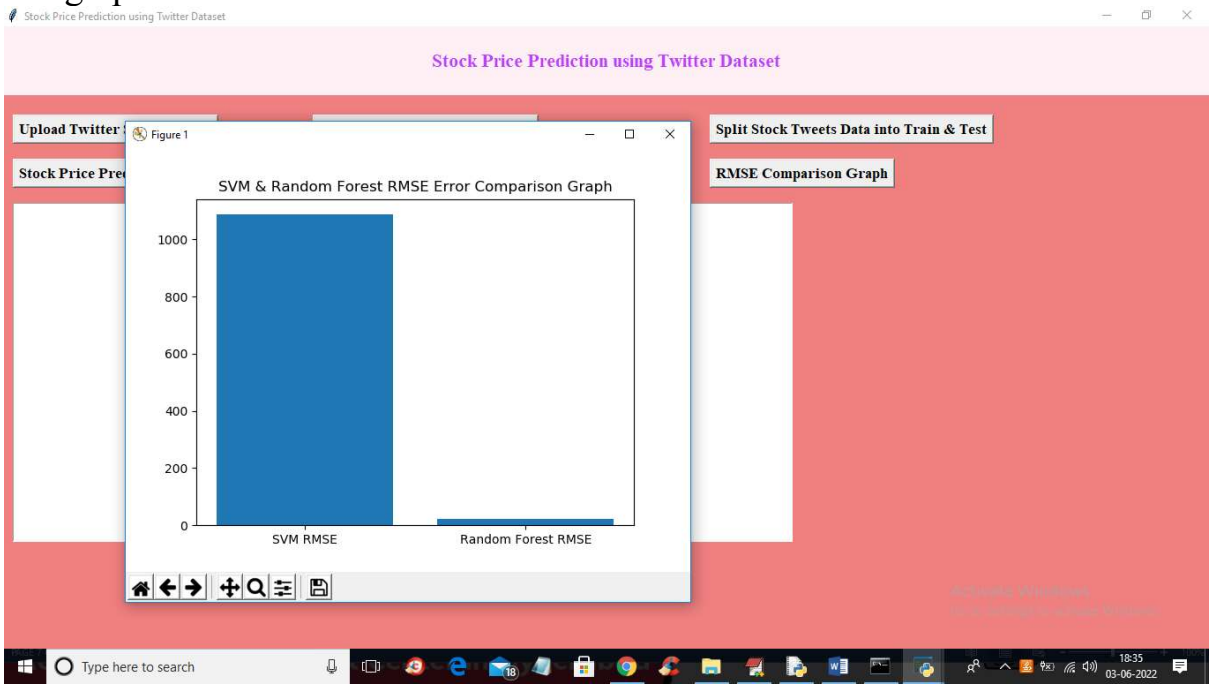
International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 5 Volume No.7 September - October – 2023
DOI:10.46647/ijetms.2023.v07i05.003 ISSN: 2581-4621

In above screen in text area we can see original test stock prices and SVM stock prices and we got SVM RMSE as 1086(difference between predicted and original prices) and in above graph x-axis contains NUMBER of DAYS and y-axis represents STOCK prices and red line represents original stock prices and green line represents SVM predicted prices and we can see both lines are overlapping with little gap and difference so we can say SVM performance is good but not best so close above graph and then click on 'Stock Price Prediction using Random Forest' button to train with Random Forest and get below output



In above screen with Random Forest we got RSME as just 21 and in graph we can see both lines are fully overlapping so TEST prices and Random Forest predicted prices are accurate. So Random Forest is best in performance

Now close above graph and then click on 'RMSE Comparison Graph' button to get below graph



DOI:10.46647/ijetms.2023.v07i05.003 ISSN: 2581-4621

In above graph x-axis represents algorithm names and y-axis represents error rate and in both algorithms Random Forest got less error so its performance is best. The lower the error the better is the algorithm

CONCLUSION :

We have investigated the causative relation between public mood as measured from a large scale collection of tweets from twitter.com and the DJIA values. Our results show that firstly public mood can indeed be captured from the large-scale Twitter feeds by means of simple natural language processing techniques, as indicated by the responses towards a variety of socio-cultural events during the year 2009. Secondly, among the observed dimensions of moods, only calmness and happiness are Granger causative of the DJIA by 3-4 days. Thirdly, a Self Organizing Fuzzy Neural Network performs very good in predicting the actual DJIA values when trained on the feature set consisting of the DJIA values, Calm mood values and Happiness dimension over the past 3 days. The performance measure we have used is kfold sequential cross validation, which is more indicative of the market movements for financial data. Finally a naive implementation of portfolio management using our strategy indicates a decent profit over a range of 40 days.

Our results are in some conjunction with [1], but there are some major differences as well. Firstly our results show a better correlation between the calm and happy mood dimensions with the DJIA values, unlike their result, which showed high correlation with only calm mood dimension. Secondly, we haven't been able to obtain high percentage result of 87%, but our 75.56% result using k-fold sequential cross validation gives stronger evidence that the correlation is over the entire range of data. The profits in out naive implementation of portfolio management shows that our MAPE estimates are pretty much accurate.

Finally, its worth mentioning that our analysis doesn't take into account many factors. Firstly, our dataset doesn't really map the real public sentiment, it only considers the twitter using, english speaking people. It's possible to obtain a higher correlation if the actual mood is studied. It may be hypothesized that people's mood indeed affect their investment decisions, hence the correlation. But in that case, there's no direct correlation between the people who invest in stocks and who use twitter more frequently, though there certainly is an indirect correlation - investment decisions of people may be affected by the moods of people around them, ie. the general public sentiment. All these remain as areas of future research.

REFERENCES :

[1] J. Bollen and H. Mao. Twitter mood as a stock market predictor. IEEE Computer, 44(10):91–94.
[2] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1–27:27, 2011.
[3] G. P. Gang Leng and T. M. McGinnity. An on-line algorithm for creating self-organizing fuzzy neural networks. Neural Networks, 17(10):1477–1493.
[4] A. Lapedes and R. Farber. Nonlinear signal processing using neural network: Prediction and system modeling. In Los Alamos National Lab Technical Report.
[5] A. E. Stefano Baccianella and F. Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In LREC. LREC.