PREDICTION ON STOCKS USING DATA MINING

Shila Jawale (Guide)

Department of Information technology

Datta Meghe College of Engineering

Airoli, India

shilaph@gmail.com

Shweta Yeshwant Nimje

Department of Information technology

Datta Meghe College of Engineering

Airoli, India

shwetanimje6@gmail.com

Ritesh Mayya

Department of Information technology

Datta Meghe College of Engineering

Airoli, India

riteshmayya@gmail.com

Mirza Nauman Ali Baig

Department of Information technology

Datta Meghe College of Engineering

Airoli, India

mailingnauman@gmail.com

***Abstract*—Stock market is a very volatile space. Accurately predicting the changes in the stock prices may prove exceedingly profitable to the investors and assist them in making smarter decisions. This research subject uses Twitter sentiment analysis to obtain the overall sentiment of the users towards the company in question which ideally leads to the changes in the stock market prices. This study attempts to implement a data mining technique called Random Forest Algorithm and use the same with the twitter sentiment score of the company to accurately predict the fluctuations in the stock market.**

Keywords—Data mining, stock, Random Forest, Twitter sentiment analysis.

I. INTRODUCTION

Predicting stock prices has been a popular topic for literature survey. Still the research is being carried out to find the best way to get money through stock market activity. Overall, the aim is to predict the future. The similar terms for prediction markets are decision markets, future ideas, virtual markets, informative markets and predictive markets[1]. Every second the market prices rise or fall that means changing constantly. Therefore, it becomes difficult to predict and invest in the market. There are different techniques determined to analyze the rise and fall of stocks. Stock means owning the shares of the company. If company ownership is divided in 100 parts and we are the investor purchasing one part which is equal to one share then we own one percent of that company [1]

Data mining is the extraction of useful and trivial patterns or knowledge from large data sets. Alternative names for data mining are knowledge discovery from data (KDD), knowledge extraction, pattern analysis, business intelligence. Whereas, plain search in goggle engine or query firing on relational database is not data mining. There are some domains of data mining such as machine learning, cognitive learning, statistics, algorithms, pattern recognition and virtualization. Files, databases and other repositories consist of huge amounts of data, hence it is necessary to develop a prevailing tool for analysis and explanation of data and extracting interesting knowledge to facilitate in decision making[2]. Some of the functionalities of data mining are the discovery of concept or class descriptions, associations and correlations classifications, prediction, clustering, trend analysis, outlier and deviation analysis, and similarity analysis [3].

Sentiment analysis is the process of determining people‘s attitudes, opinions, evaluations, appraisals and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes[4]. Basically, it is the one’s judgment or evaluation on some topic or the polarity of the document. A basic job of sentimental analysis is to collect all required data that may be a single sentence, a whole paragraph or line from respective tweets and analyzing its positivity and negativity for a better result.

II. PROBLEM DEFINITION

Stock markets are incredibly large and hard to grasp its behavior. There are too much of variations present in the result of stocks. People’s main aim in stock market is to make profit by buying or selling the stocks, but due to many ups and downs in the stock price with respect to time it become difficult to go with the stocks. Thus there is a necessary in prediction of stocks. But due to this large market volatility it is considered too unpredictable to be reliable. Values of Stock market is varied due to many aspects such as Historical Data, Tweets, News, Reputation of that company, natural calamities, global financial disturb and many more.

Funding in a strong stock but at a bad moment may have catastrophic consequences; at the same moment investing at a good time will produce better income. Stock holders face this trading issue because they don't fully grasp which stocks to purchase at a particular time or which stocks to sell to get effective outcomes. So we tried to overcome this problem by using regressor algorithm and twitter sentiment analysis to predict up to its extent.

III. PROPOSED SYSTEM

The solution proposed from this project is to use Twitter sentiment analysis to predict the rise or fall of the price of a stock. This is done by fetching raw historical data of the stock along with most recent tweets related to that company. These tweets are analyzed using text mining. For example along with the name of the company the words used are good, great or any other positive words then the result is positive and the result is negative otherwise. This information is processed using the Random forest algorithm. After which we get many features along with a positive and a negative feature. These positive and negative features are selected and classified so that we can get the overall result. It may be positive or negative. This helps the investor make an intelligent decision. For our proposed system we at first collected the historical data from the internet via the Yahoo! Finance, it provides the original content of financial reports, useful financial historical data, stock data. We used python language for our problem statement, python has a library named yfinance by which it is reliable to download the historical data of stocks of a particular company. Further tweets are retrieved through the API of twitter named tweepy, this would easily able to retrieve whole information of about a particular tweets for examples, tweets, ID of a user who had tweeted date and time of tweet, location, likes and retweets for that tweets, etc. Thus by applying sentiment analysis over tweets results into the sentiment values combining these values and the result obtained from the algorithm applied to historical data will conclude the prediction. Figure 1 shows the flow of our system.

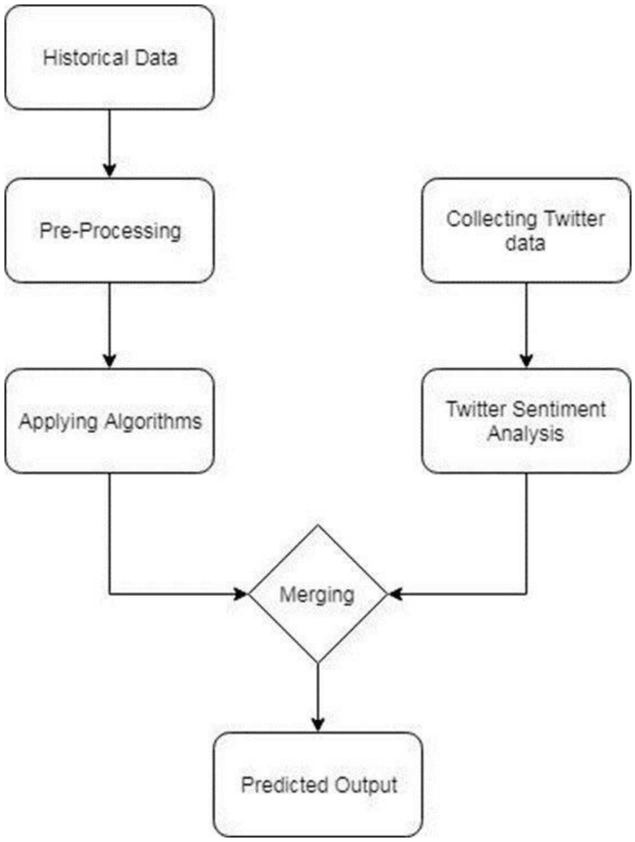


Fig 1. System Architecture

IV. METHODOLOGIES

A. *Random Forest Algorithm*

Random forest algorithm is used for classification as well as regression. It is used for stock market prediction. This algorithm is flexible to handle missing values as well as it won't overfit the model. As stock prices are volatile in nature, predicting is quite challenging. As the name suggests, this algorithm creates the forest with different parameters that is the number of trees. The algorithm works by selecting random samples from a given dataset. Next, it will construct a decision tree for every sample. Then it will get the prediction result from every decision tree. After that, voting for every predicted result will be evaluated. At last, most voted predicted result will be the predicted output. The dataset we used from the Yahoo Finance, 80% of data was used to train the machine according to our model and 20% to test the data. Thus, the basic approach is to learn the patterns and relationships from the training set and reproduce them to the test data.

B.Twitter Sentiment Analysis

Consumers usually express their sentiments on public forums like social network sites like Facebook and Twitter. The data collected from real time would result in accurate results for prediction of stocks. Opinions, feelings, comments are all in slang or disorganized manner. Manual analysis of such data is virtually impossible. Python library and Tweepy allows text preparation, sentiment detection, sentiment classification, and at last presentation of output. Specifically, it eliminates the irrelevant data and extracts the text relevant to the area of study from the data. Sentimental analysis is done at different levels such as negation, lemmas etc. Sentimental classification is the next important step where groups are classified into good, bad, positive, negative, like, dislike. Python allows you to represent the data using line graph, bar chart, pie chart.

V. MODULE

IDENTIFICATION *A. Data Collection*

Data collection is the initial step of any project. The right collection of data is the important aspect. The data collected are never ready to implement any algorithm. Collecting data for the relevant project will make the process easy. We have collected data from NSE websites of different companies. Initially, we will be analyzing the dataset according to the model and predict the results accurately.

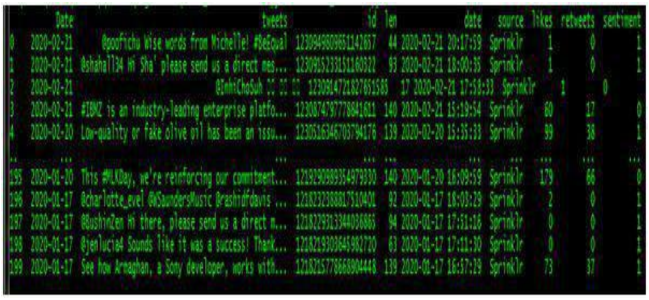
B.Pre Processing

Data preprocessing means converting raw data into efficient and useful data. Different processes involve data cleaning, data transformation, data reduction. In data cleaning missing, noisy data are eliminated. Next data transformation where data normalization, discretization actions are carried. Data reduction aims to reduce the storage efficiency and reduce data storage

C.Training the machine/data score

In Data mining process training the data plays an important role so as to get an accurate result of our prediction. For the algorithm we used the dataset of stock market containing the parameters like Date, Open price, Close price, High price, Low price, Adjacent close and Volume. Each single dataset belongs to a particular company. We have used Yahoo! Finance market data downloader to retrieve the data; "yfinance" aims to offer a reliable, threaded, and Pythonic way to download historical market data from Yahoo! finance.

Twitter is a popular social network where users share messages called tweets. Twitter allows us to mine the data of any user using Twitter API or Tweepy. The data will be tweets extracted from the user. The first thing to do is get the consumer key, consumer secret, access key and access secret from twitter developer available easily for each user. These keys will help the API for authentication. Tweepy is one of the libraries that can be installed using pip. Tweepy provides the convenient Cursor interface to iterate through different types of objects. Thus we could retrieve the tweets related to provided keywords of the company; it includes used_id, Tweets, date, time, retweets, likes, along with Sentiments.



```
          Date                                          tweets                   id  len                 date   source  likes  retweets  sentiment
0   2020-02-21        @goofichu Wise words from Michelle! #BeEqual  1230949800965114267   44  2020-02-21 20:17:59  Sprinklr      1         0          1
1   2020-02-21  @shahall34 Hi Sha' please send us a direct mes...  1230915233151160322   93  2020-02-21 18:00:35  Sprinklr      1         0          1
2   2020-02-21                              @InhiChoSuh     1230914721827651585   17  2020-02-21 17:58:33  Sprinklr      1         0
3   2020-02-21  #IBMZ is an industry-leading enterprise platfo...  1230874797778841611  140  2020-02-21 15:19:54  Sprinklr     60        17          0
4   2020-02-20  Low-quality or fake olive oil has been an issu...  1230516346703794176  139  2020-02-20 15:35:33  Sprinklr     99        38          1
..         ...                                               ...                  ...  ...                  ...      ...    ...       ...        ...
195 2020-01-20  This #MLKDay, we're reinforcing our commitment...  1219290989354979330  140  2020-01-20 16:09:59  Sprinklr    179        66          0
196 2020-01-17  @charlotte_evel @NSaundersMusic @rashidfdavis ...  1218232388017510401   92  2020-01-17 18:03:29  Sprinklr      2         0          1
197 2020-01-17  @BushinZen Hi there, please send us a direct m...  1218229313344036865   94  2020-01-17 17:51:16  Sprinklr      0         0          1
198 2020-01-17  @jenlucia4 Sounds like it was a success! Thank...  1218219303645982720   63  2020-01-17 17:11:30  Sprinklr      0         0          1
199 2020-01-17  See how Armaghan, a Sony developer, works with...  1218215778668904448  139  2020-01-17 16:37:29  Sprinklr     73        37          1
```

Fig. 2 Sentiment Scores of Tweets

VI. EXPERIMENTAL RESULTS

For our stocks, we used the Random Forest Algorithm; Random Forests are based on learning techniques for the ensemble. Ensemble means literally a group or set, which in this case is a set of decision trees, called a random forest together. The reliability of ensemble models is higher than the accuracy of individual models as it compiles the results from the individual models and produces a final result. Properties are automatically chosen using a process known as bootstrap averaging or bagging. From the set of features available in the dataset, a number of training subsets are created by choosing random features with replacement. What this means is that one feature may be repeated in different training subsets at the same time.

In the training data set, stocks are divided into N classes based on the forward excess returns of each stock. The trained RF model is then used in the subsequent trading period to predict the probability for each stock. We construct our random forest model with no change in it. No modification is made to the algorithm, as it is believed that the original RF can have enough capacity to handle large numbers of variables in datasets and give rise to unbiased estimates for real world classification problems, including finance.

In principle, the random forest consists of many deep but uncorrelated decision trees built upon different samples of the data. The process of constructing a random forest is simple. For each decision tree, we first randomly generate a subset as a sample from the original dataset. Then, we grow a decision tree with this sample to its maximum depth of ‘Sd’. Meanwhile, ‘sp’ features used on each d split are selected at random from ‘p’ features. After repeating the procedure numerous times with the original dataset, ‘O’ decision trees are generated. The final doutput is an ensemble of all decision trees, and the classification is conducted via a majority vote. The computational complexity can be simply estimated as

$$O(O(p*nins*lognins)) \qquad (1)$$

Where d ‘nins’ represents the number of instances in the training datasets. Three parameters must be tuned to check the robustness of the RF on classification, i.e., the number of trees O,the maximum d depth S and the number of features spd of each split. We set the d maximum depth S to be unlimited so that the nodes are expanded until all leaves are pure or until all leaves contain less than two samples. Regarding the feature sub sampling, we typically choose sp=√p. The influence d

Of the number of trees on the classification accuracy and the out-of-sample performance is then systematically investigated.

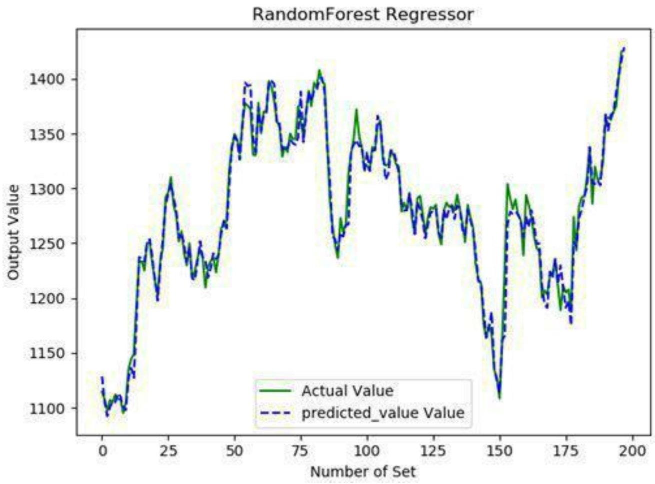


Fig. 3 Output of Random Forest Algorithm

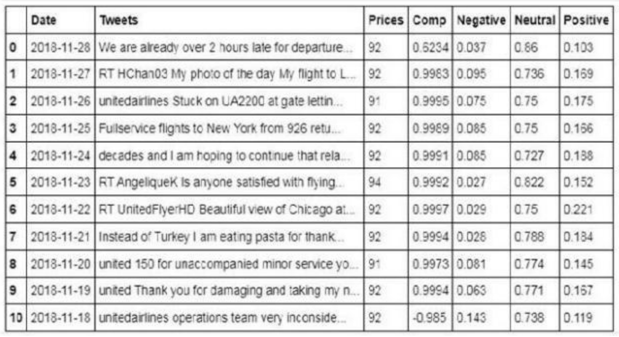
Figure 3, shows the output of a regression model with trained dataset and test dataset of stocks this model gives an accuracy around 85%, thus increasing the accuracy we approached to twitter sentiment analysis. 700-900 data were collected as historical data of stocks. Tweeter data were also collected for the same period of time. The Twitter data is available for all days lying in the giving period, the stock 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 So if the stock value on a given day is x and the next accessible data point is y with n days left in between, we estimate the missed data by calculating to all be (y+x)/2 on the first day after x and then continuously using the same approach before all the holes are filled.

At first we retrieve the tweets and try to clean the tweets, cleaning the tweets include removing all hash tags, unnecessary, spaces and tabs, and all special character.

Further applying sentiment analysis over tweets we get the respective sentiment scores, these scores appears as a percentage of the obtaining result i.e. the positive and negative result of tweets. When tweets were collected and their polarity is decided, the next step was to collect data from the stock exchange market. Data was collected via Yahoo finance. We have considered closing the price column as our target, thus we clubbed the tweets from twitter, and price from stocks on that particular date. Figure 2 shows the dataset along with sentiment values of a tweets (here we have taken an example of TCS company)

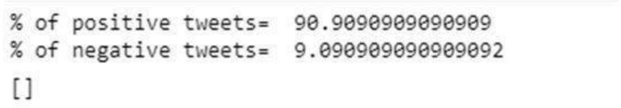
TABLE 1

Sentiment values



	Date	Tweets	Prices	Comp	Negative	Neutral	Positive
0	2018-11-28	We are already over 2 hours late for departure...	92	0.6234	0.037	0.86	0.103
1	2018-11-27	RT HChan03 My photo of the day My flight to L...	92	0.9983	0.095	0.736	0.169
2	2018-11-26	unitedairlines Stuck on UA2200 at gate lettin...	91	0.9995	0.075	0.75	0.175
3	2018-11-25	Fullservice flights to New York from 926 retu...	92	0.9989	0.085	0.75	0.166
4	2018-11-24	decades and I am hoping to continue that rela...	92	0.9991	0.085	0.727	0.188
5	2018-11-23	RT AngeliqueK Is anyone satisfied with flying...	94	0.9992	0.027	0.822	0.152
6	2018-11-22	RT UnitedFlyerHD Beautiful view of Chicago at...	92	0.9997	0.029	0.75	0.221
7	2018-11-21	Instead of Turkey I am eating pasta for thank...	92	0.9994	0.028	0.788	0.184
8	2018-11-20	united 150 for unaccompanied minor service yo...	91	0.9973	0.081	0.774	0.145
9	2018-11-19	united Thank you for damaging and taking my n...	92	0.9994	0.063	0.771	0.167
10	2018-11-18	unitedairlines operations team very inconside...	92	-0.985	0.143	0.738	0.119

Further analyzing our data we arrived at a result figure 5 showing a high percentage of Positive tweets resulting into rise in the stock price of that company.



```
% of positive tweets=  90.9090909090909
% of negative tweets=  9.090909090909092
[]
```

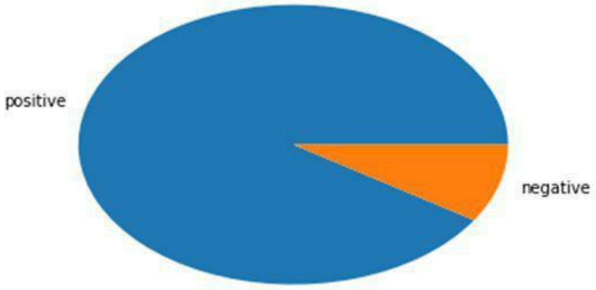


Fig. 4 Pie Chart of Sentiments

VII. SCOPE OF THE PROJECT

We have investigated the causative relation between User sentiments as measured from a large scale collection of tweets from twitter.com and the stock values. Our results show that firstly public mood can indeed be captured from the large-scale. Twitter feeds by means of simple Sentiment analysis. Our results are in some conjunction, but there are some major differences as well. Firstly, our results show a better correlation between the positive, negative, and neutral dimensions with the NSE values, unlike other, which showed high correlation with only neutral mood dimension.

In a potential course, work would like to test and apply a model of economic growth for stock market prediction and examine how economic growth models can impact stock market prediction. this work would like to conduct a comparative study of deep learning classifiers and severe learning classifiers centered on the parameters used for stock market modeling using a feature reduction algorithm.

It's possible to obtain a higher correlation if the actual mood is studied. It may be hypothesized that people's mood indeed affects their investment decisions, hence the correlation.

VIII. Conclusion

The solution proposed in this paper is to use twitter sentiment analysis to predict the rise or fall of the price of a stock. This is done by fetching raw historical data of the stock along with most recent tweets related to that company. These tweets are analyzed using text mining. For example along with the name of the company the words used are good, great or any other positive words then the result is positive and the result is negative otherwise. This information is processed using the Random Forest algorithm. After which we get many features along with a positive and a negative feature. These positive and negative features are selected and classified so that we can get the overall result. It may be positive or negative. This helps the investor make an intelligent decision.

IX. References

[1] Kute, Shyam, and Sunil Tamhankar. "A survey on stock market prediction techniques." *International Journal of Science andResearch* (2013)

[2] Khedr, Ayman E., and Nagwa Yaseen. "Predicting stock market behaviour using data mining technique and news sentiment analysis." *International Journal of Intelligent Systems andApplications* 9.7 (2017): 22.

[3] Maini, Sahaj Singh, and K. Govinda. "Stock market prediction using data mining techniques." *2017 International Conference on Intelligent Sustainable Systems (ICISS)*.IEEE, 2017.

[4] Sharma, Ashish, Dinesh Bhuriya, and Upendra Singh. "Survey of stock market prediction using machine learning approach*2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*.Vol. 2 IEEE, 2017.

[5] Alostad, Hana, and Hasan Davulcu. "Directional prediction of stock prices using breaking news on Twitter." *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent IEEE, 2015 Smart City and Emerging Technology (ICSCET)*.IEEE, 2018.

[6] Mankar, Tejas, et al. "Stock Market Prediction based on Social Sentiments using Machine Learning." *2018 International Conference on Smart City and Emerging Technology (ICSCET)*. IEEE, 2018.

[7] Navale, G. S., et al. "Prediction of stock market using data mining and artificial intelligence." *International Journal of Engineering Science 6539 (2016).*