PREDICTION ON STOCKS USING DATA MINING

Shila Jawale (Guide)
Department of Information technology
Datta Meghe College of Engineering
Airoli, India
shilaph@gmail.com

Shweta 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

Nauman Ali Mirza
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 may prove exceedingly profitable to the investors and assist them in making smarter decisions. This can be made possible by using various data mining techniques to predict the fluctuations in the stock market. This study attempts to implement different data mining techniques to choose the most suitable method.**

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 idea, 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 google 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 judgement or evaluation on some topic or the polarity of the document. A basic job of sentimental analysis is grouping the extremity of a given content at the record, sentence, or associate degree angle level - no matter whether or not the communicated feeling in an archive, a sentence or a substance include/perspective is bound, negative, or impartial

II. PROBLEM DEFINITION

Stock market is very vast and difficult to understand. It is considered too uncertain to be predictable due to huge fluctuation of the market. Stock market prediction tasks are interesting as well as divides researchers and academics into two groups, those who believe that we can devise mechanisms to predict the market and those who believe that the market is efficient and whenever new information comes up the market absorbs it by correcting itself, thus there is no space for prediction.

Investing in a good stock but at a bad time can have disastrous results, while investing in a stock at the right time can bear profits. Financial investors of today are facing this problem of trading as they do not properly understand as to which stocks to buy or which stocks to sell in order to get optimum results. So, the proposed project will reduce the problem with suitable accuracy faced in such a real time scenario.

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. The prime aim of the proposed system is to fetch live server data by using the Python programming language, which can be used for performing sentiment analysis on the extracted datasets from Twitter. In this context, first python is installed on the host machine, after that required software/API such as tweepy is installed using terminal command prompt environment. For Debian or Ubuntu platform tweepy can be installed with system package manager. Tweepy uses Python library for fetching live data, and this tool helps to pull contents from desired web pages then save the required information. For illustration purposes we fetch the live data of Sensex and Nifty that can be used further for sentiment analysis.

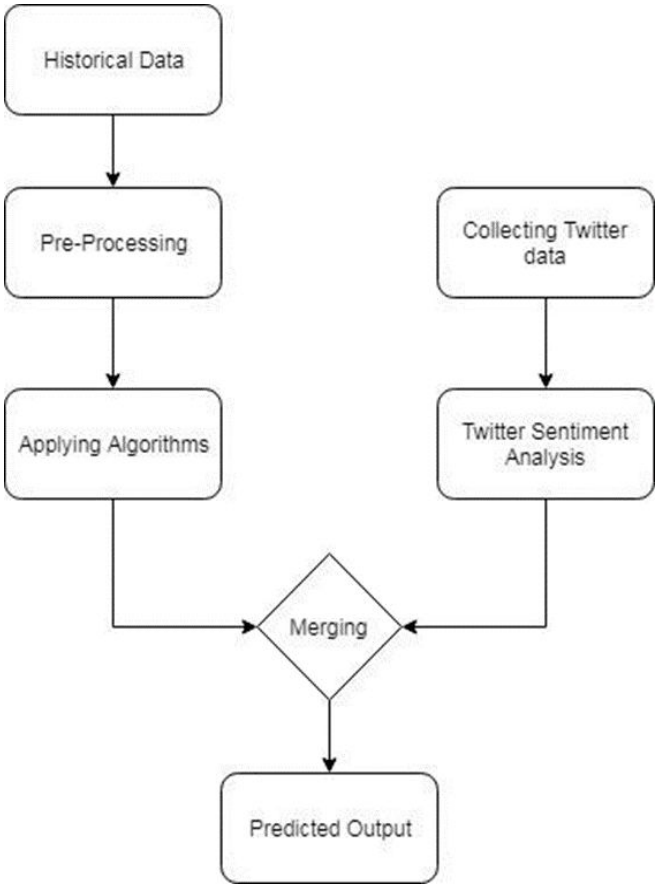


Fig 1 System Architecture

IV. METHODOLOGIES

1.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 data set we used from the year 2014 from the NSE ,80% of data was used to train the machine and rest 20% to test the data. Thus, the basic approach is to learn the patterns and relationship from the training set and reproduce them to the test data.

2.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 extract 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

1.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.

2.Pre Processing

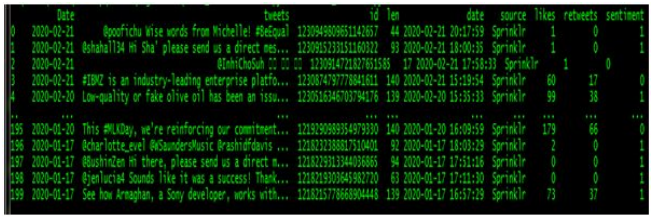
Data preprocessing means converting raw data into efficient and useful data. Different processes involve data cleaning, data transformation, data reduction. In data cleaning missing, noisy data are eliminated. Next data transformation where data

normalization, discretization actions are carried. Data reduction aims to reduce the storage efficiency and reduce data storage.

3. 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 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. To authorize our app to access Twitter on our behalf, we used the OAuth Interface. 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	@poofichu Wise words from Michelle! #BeEqual	1230948009651142657	44	2020-02-21 20:17:59	Sprinklr	1	0	1
1	2020-02-21	@ohahaTl34 Hi Sha' please send us a direct mes...	1230915233151160322	93	2020-02-21 18:00:35	Sprinklr	1	0	1
2	2020-02-21	@AnhChoSuh 🙏🙏🙏	1230914721827651585	17	2020-02-21 17:58:33	Sprinklr	1	0	
3	2020-02-21	#IBMZ is an industry-leading enterprise platfo...	1230874797778841011	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...	1219390888354979330	140	2020-01-20 16:00:59	Sprinklr	179	66	0
196	2020-01-17	@charlotte_evel @WSaundersMusic @rashidfdavis ...	1218392388817510401	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 17:11:16	Sprinklr	0	0	1
198	2020-01-17	@jenlucia4 Sounds like it was a success! Thank...	1218219303645982720	63	2020-01-17 17:11:10	Sprinklr	0	0	1
199	2020-01-17	See how Armaghan, a Sony developer, works with...	1218215778668904448	139	2020-01-17 16:57:29	Sprinklr	75	37	1

Fig 2

V. EXPERIMENTAL RESULTS

We have used the Random forest algorithm for our stocks, Random forests are based on ensemble learning techniques. Ensemble, simply means a group or a collection, which in this case, is a collection of decision trees, together called a random forest. The accuracy of ensemble models is better than the accuracy of individual models due to the fact that it compiles the results from the individual models and provides a final outcome. Features are selected randomly using a method known as bootstrap aggregating 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 'S_d'. Meanwhile, 'sp_d' features used on each split are selected at random from 'p' features. After repeating the procedure numerous times with the original dataset, 'O_d' decision trees are generated. The final output is an ensemble of all decision trees, and the classification is conducted via a majority vote. The computational complexity can be simply estimated as

$$O(O_d(p*nins*lognins)), \text{ where}$$

'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_d, the maximum depth S_d and the number of features sp_d of each split. We set the maximum depth S_d 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 subsampling, we typically choose $sp_d=\sqrt{p}$. The influence of the number of trees on the classification accuracy and the out-of-sample performance is then systematically investigated.

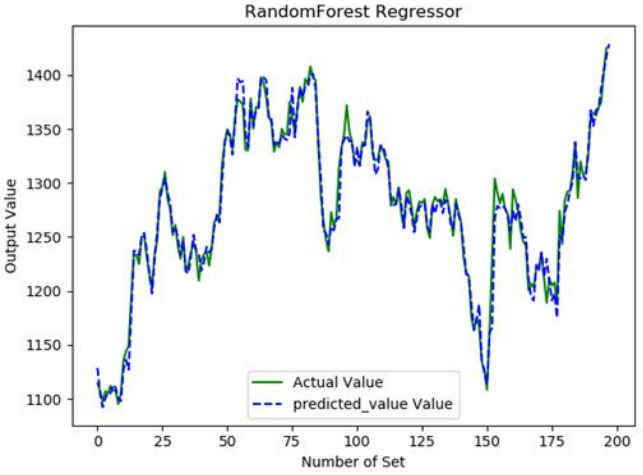


Fig 3

figure 3, shows the output of regression model with trained dataset and test dataset of stocks thi model gives an accuracy around 85%, thus increasing the accuracy we approached to twitter sentiment analysis.

Stocks were collected for over the period from february 1st, 2017 to february 1st , 2020 from stocks data. In total, around 1000 data were collected. Tweets were 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 using a concave function. So, if the stock 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.

First step is preprocessing Twitter data. To decide polarity of tweets and by polarity we mean decision weather tweet is positive, negative or neutral. Tweets that have score smaller than 0 is decided to be negative, for the ones that have score higher than 0 was decided to be negative and the ones that have score 0 have neural polarity. For the In Table 1 is example of tweets related to Microsoft: Table1. Samples of collected tweets and their scores Text of tweet. When tweets were collected and their polarity decided, next step was to collect data from stock exchange market. Data was collected via Yahoo finance. we have considered closing price column as our target, thus we clubbed the tweets from twitter, and price from stocks on that particular date. figure 4 shows the dataset along with sentiment values of a tweets (here we have taken an example of TCS company)

	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

Fig 4

further analyzing our data we arrived at a result figure 5 showing 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
[]
```

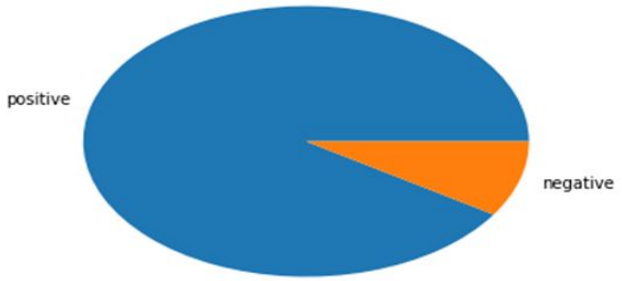


Fig 5

VI. 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 NSE 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.

As a future direction, this research would like to perform a comparative analysis with deep learning classifiers and extreme learning classifiers with the help of a feature reduction algorithm based on the parameters used for stock market prediction. Along with this, research would also like to study and implement an economic growth model for stock market prediction and the analysis of how economic growth model will affect in stock market prediction in comparison to the linear regression model and with specialized machine learning techniques. 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. Fig.1 Proposed Model for Predicting stock market using sentiment analysis.

VII. 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.

VII. References

1. Kute, Shyam, and Sunil Tamhankar. "A survey on stock market prediction techniques." *International Journal of Science and Research* (2013)
2. Das, Debashish, and Mohammad Shorif Uddin. "Data mining and Neural network techniques in Stock market prediction: A methodological review." *International journal of artificial intelligence & applications* 4.1 (2013): 117.
3. Al-Radaideh, Qasem A., Adel Abu Assaf, and Eman Alnagi. "Predicting stock prices using data mining techniques." *The International Arab Conference on Information Technology (ACIT'2013)*. 2013.
4. Khedr, Ayman E., and Nagwa Yaseen. "Predicting stock market behavior using data mining technique and news sentiment analysis." *International Journal of Intelligent Systems and Applications* 9.7 (2017): 22.
5. Mittal, Anshul, and Arpit Goel. "Stock prediction using twitter sentiment analysis." *Stanford University, CS229 (2011 http://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis. pdf)* 15 (2012).
6. Maini, Sahaj Singh, and K. Govinda. "Stock market prediction using data mining techniques." *2017 International Conference on Intelligent Sustainable Systems (ICISS)*. IEEE, 2017.
7. 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.
8. 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 Technology (WI-IAT)*. Vol.

1. IEEE, 2015 *Smart City and Emerging Technology (ICSCET)*. IEEE, 2018.

9. 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.
10. Navale, G. S., et al. "Prediction of stock market using data mining and artificial intelligence." *International Journal of Engineering Science* 6539 (2016).
11. Si, Jianfeng, et al. "Exploiting topic based twitter sentiment for stock prediction." *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vol. 2. 2011
12. Oh, Chong, and Olivia Sheng. "Investigating Predictive Power of Stock Micro Blog Sentiment in Forecasting Future Stock Price Directional Movement." *Icis*. 2011.