International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

Applying of a Company's Stock Price Prediction using Data Mining

S.Muthuselvi, A.Rengarajan, S.Scinthiaclarinda, K.Nithya

Abstract*: Stock market analysis is a common economic activity that has been an attractive topic to research and used in different forms of day-to-day life in order to predict the stock prices. Techniques like major analysis, Statistical investigation, Time arrangement analysis and so on are reliably worthy forecast device. In this paper, Data mining, Machine learning (ML) and Sentiment analysis are techniques used for analyzing public emotions in order predict the future stock prices. The goal of a project is to review totally different techniques to predict stock worth movement victimization the sentiment analysis from social media, data processing. Sentiment classifiers are designed for social media text like product reviews, blog posts, and email corpus messages. In the company's communication network, information mining calculation is utilized as to mine email correspondence records and verifiable stock costs. Implementing various Machine learning and Classification models such as Deep Neural network, Random forests, Support Vector Machine, the company can successfully implemented a company-specific model capable of predicting stock price movement with efficient accuracy.

Keywords: Data Mining, Machine Learning, Random Forest, Deep Neural Network, Support Vector Machine.

I. INTRODUCTION

Securities exchange expectation is chiefly used to decide the future estimation of here and there's of financial exchange cost. The forecast of securities exchanges is viewed as a difficult errand of monetary time, arrangement expectation [4]. Securities exchange forecast is trying essentially on account of the unconventionality of the financial exchange with its loud and unstable condition, considering the solid association with various stochastic factors, for example, political occasions, papers just as quarterly and yearly reports [3]. Social media offers a robust outlet for people thoughts and feelings. News and updates are erratic and it is extraordinary enthusiasm to assess if there is a connection between an association's stock and the open feeling [6]. For this, the one methodology is to break down the open feeling of an association so as to figure the advancement of the association's stock. Examination of internet based life is unequivocally identified with assessment investigation [1] which is normally utilized in numerous ventures and gives partners incredible devices for seeing how the individual responds to specific occasions. This is utilized to remove feelings and sentiments from content. Data mining (DM) is the way toward dealing with enormous informational collections to recognize designs and build up connections to take care of issues through information examination. Information mining is the way toward extricating the helpful information, examples and patterns from a lot of dataset. Information mining [3] is an interdisciplinary subfield of software engineering with a general target to separate information from an informational collection and change the information into a possible structure for further use. Close to the rough examination step, it in like manner incorporates database and the board perspectives, information pre-handling, model and surmising contemplations.

Enron corpus dataset was gathered and arranged by the CALO Project (A Cognitive Assistant that learns and Organizes). The email dataset contains information from around 150 clients, for the most part senior administration of Enron, composed into envelopes. It contains a sum of about 0.5M messages. Invalid email delivers [11] were changed over to something of the structure user@enron.com at whatever point conceivable (i.e., recipient is specified in some parse-able format like "Doe, John" or "Mary K. Smith"). Enron email corpus dataset is used to research the issues in analyze the public emotion of an organization to stock market prediction. Company's stock price performance can predicted by internal communication pattern. Classifiers like Random Forest and Deep Neural networks, Support Vector Machine are used for predicting stock price movements. The remaining of this paper is as follows. Section 2 introduces the background of proposed work and provides a survey of related literature survey to the identified problem. Section 3 stated and explains our proposed methods to handle the research problem. Lastly, Section 4 describes the paper with the future ideas.

II. RELATED WORKS

Examination of financial exchange expectation is the demonstration of deciding the future estimation of an organization's stock cost.

Revised Manuscript Received on 30 July 2019.

***** Correspondence Author

S.Muthuselvi*, PG Scholar, Computer Science and Engineering, Vel Tech Multi Tech Dr.RangarajanDr.Sakunthala Engineering College, Chennai, India.

A.Rengarajan, Professor, Computer Science and Engineering, Vel Tech Multi Tech Dr.RangarajanDr.Sakunthala Engineering College, Chennai, India.

S,Scintiaclarinda, PG Scholar, Computer Science and Engineering, Vel Tech Multi Tech Dr.RangarajanDr.Sakunthala Engineering College, Chennai, India.

K.Nithya, Assistant Professor, Computer Science and Engineering, Vel Tech Multi Tech Dr.RangarajanDr.Sakunthala Engineering College, Chennai, India.

Securities exchange expectation assumes a significant job in the development of the organization. Different approaches to examination the securities exchange forecast utilizing information mining are given beneath. As of late, there has been put some push to dig web-based social networking for open conclusion examination [12]. Studies have proposed that open feelings appeared through speaker unit could well be connected with the Dow-Jones Industrial Average Industrial Average. Amid this article, it will in general propose a method to mine Twitter learning. In particular, we propose to utilize a data mining algorithmic program to check whether the cost of a scope of thirty firms recorded in information framework and along these lines the New York securities market will genuinely be predicted by the given fifteen million records of tweets (i.e., Twitter messages).Thus by extricating vague issue tweet learning through IP strategies to diagram open conclusion, at that point manufacture utilization of a data mining procedure to get designs between open feeling and genuine stock esteem movements[7]. With the anticipated calculation, it will in general figure out how to get that it's achievable for the stock estimation of certain organizations to be predicted with a middle exactness as high as seventy six.12%. The system proposes for developing the mechanical stock exhibitions [2]. The expectation of stock value utilizing information mining strategies connected to specialized factors has been generally looked into however very little research to information has been done in applying information mining methods to both specialized and central data. In this, a structure is built that empowers us to make class forecasts about modern stock exhibitions. So as to have a systemized methodology for the choice of stocks and expanding stock value execution, a few diagnostic procedures are connected. Two objectives are here to approve our stock choice philosophy and to decide if our exchanging methodology enables us to beat the Australian market. These recreated outcomes demonstrate that our chose stock portfolios beat the Australian All-Ordinaries Index. The finding of our system legitimizes the utilization of investigation for grouping and expectation reason. The focal point of a task is to examine whether stocks chosen by the utilization of an individual precise methodology with investigative procedures. The various information digging procedures are utilized for exploring money related information examination. Money related information investigation is utilized in numerous budgetary organizations for precise examination of purchaser information to discover defaulter and legitimate customer[3].

The analysis focuses mainly on combining forecast from different models to perform better time series. Forecasting is a method for evaluating future parts of a business [4] .Forecasts is vital for present moment and long term goals.Joining gauge from different models has seemed to perform better to single estimate in most time arrangement. The focal point of the venture is on how the best models of one arrangement can be connected to comparable recurrence design arrangement for anticipating utilizing affiliation mining [8]. By contrasting and proposed technique, results demonstrate that Holt winter strategies are better.

The above mentioned systems are focused on smaller dataset; therefore accuracy of a stock price gets decreased. Here stock price is unpredictable, when the traditional classifier is used. In this paper, larger dataset like Enron email corpus is used for analyzing the company's stock price.

III. PROPOSED METHODOLOGY

The proposed technique relies on analyzing different classifiers for stock price prediction within a company's specific business. Opinions of many people ideas are taken in a dataset. Enron email corpus is publically available dataset, which is used to extract information. The modules are:

- Data Collection
- Preprocessing
- Classifiers
  - Random Forests
  - Deep Neural Network
  - Support Vector Machine
- Result with Analysis

A. Architecture

In the Proposed system architecture, the emails in their raw form contain a lot of information that is unsuitable for modeling analysis. Therefore, the data has to be preprocessed. Then the processed dataset is then classified with different classifiers like Random forest, Support Vector Machine, Deep Neural Network [5]. Then, it is compared and visualized as a graph format using Anaconda navigator tool.

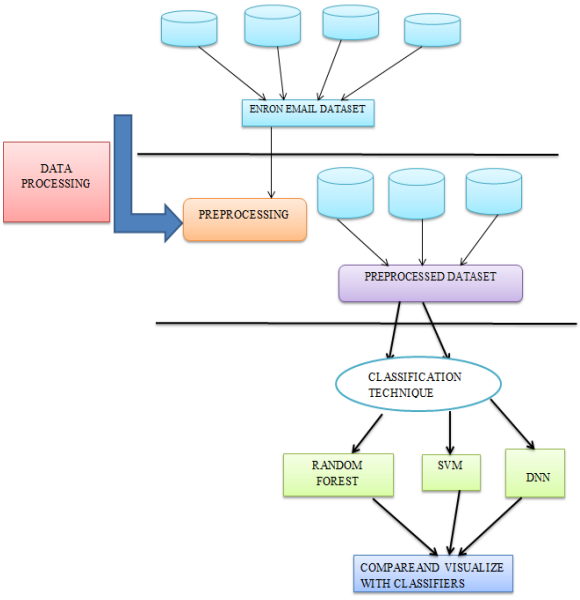


Fig 1.System Architecture of Stock Price

. Data Collection

Our dataset consists of mainly daily mail about the stock market. Each email in the Enron corpus contains the mail address of the sender and receiver(s), time, date, subject and email body. Email connections were not made accessible.

C.Pre-Processing

Pre-processing is used for standardize the dataset.The words are a mixture of extra punctuations, misspelled words and words with many repeated letters[9].

These noisy words will interfere with our learning algorithm. If we include these words then our learning algorithm seems to look for quantifiers like “a”, “an”, “this” etc.The following list of things we had to do in order to pre-process our data:

1. Delete all repeated hyper-text links from the Email data.

2. Change all word blocks from the email corpus data to lower case. This increases 14 uniformity and changes helps us to remove repetitions if present.

3. Removing white spaces from the email data. To keep the emoticons, because they provide helpful insights about the mail.

4. We remove punctuations marks like commas, full stop etc.

5. Clear out any tag to any person in the email data. Tags starting with“@” are removed. We keep hashtags because sometimes tags like “#Stock #Crash” helps us better understand the mood.

Pre-processing involved in Stock Prediction:

1. Data Cleaning:

- In Data mining, the missing values can be handled by using Pandas.
- Pandas provide the various methods for cleaning the missing values.

2. Data Reduction:

Process of reduced representation in volume but it produces the same or similar analytical results.

- Capitalization:

  Content for the most part has an alternate of capitalization mirroring the introduction of the line. The most widely recognized methodology is to lessen everything to bring down case for straightforwardness yet recall that a few words

  Precedent: "US" to "us", can change the word when it adjusted to the lower case.

- Stop word:

  The principle parts of the words in a given substance are associating portions of a sentence other than appearing, articles or expectation. Word like "the" or "and" taxi be evacuated by contrasting content with a rundown of stop word.

3. Data Wrangling:

- Counting Valid Data

  There are many null data in the Enron dataset. In order to select the most appropriate features to explore, to present at least in 70% of the dataset. Considering there are 21 attributes (from which 70% is approximate to 15 attributes), and it will initially see which examples have in excess of 15 not invalid values and choose the most total attributes from this selection.

D.Classifiers

The first step in the machine learning process is to split our data into training and testing set. The goal is to fit a model with its default parameters and predict the outcome variable (label) of POI. This is a binary classifier and will rely on supervised learning algorithms.

a) Random Forest

Random Forest (RF) is a collection of the tree predictors such that each tree depends on the values of individually sampled random vectors.The random forest uses an ensemble method to create many decision trees based on the training data. These decision trees are then averaged and an optimization of those values is used to decide which tree to use.

Algorithm Random Forest

Precondition: A training set S:= (X_1,Y_1), ….. ,(X_n,Y_n), features F, and number of trees in forest B.

1**: function** RANDOMFORESTS, F)
2: H←ϕ
3: **for**i € 1,….,B**do**
4:$S^{(i)}$←A bootstrap sample from S
5: h_i ← RANDOMIZEDTREELEARN($S^{(i)}$,F)
6: H←H U{ h_i }
7: **end for**
8: **return H**
9**: end function**
10: **function** RANDOMIZEDTREELEARN(S,F)
11: At each node:
12: f ← very small subset of F
13: Split on best feature in f
14: **return**thelearned tree
15: end function

a) Deep Neural Network

Stock market prediction takes use of DeepNeuralNetworkdue to the non-linear nature of stock price [10].The Deep Neural Network algorithm is an Artificial Neural Network, where it consists of neurons or units arranged in multiple layers, which convert an input vector into some output. In this network, each layer of nodes trains on a distinct set of features based on the previous layer’s output.

Algorithm Deep Neural Network

Input: x=$(x_1.x_2,…..x_n)$ denotes the data matrix of n samples , y=$(y_1,y_2,….y_n)^K$ as corresponding output labels, the maximum number of selected attributes A.

1. **Initialize:** F={bias}, S=R &W_s=0
2. **While**$|F| \leq t + 1$ do
3. Assign W_s=0;
4. Update weight of hidden layers as well as input weight W_s;

5. Multiple times drop out to be used and then obtain average G_{RS}
6. Calculate$j = arg \max_{f \in F} ||G_{RS}|| q$
7. Update Learning rates using AdaDelta
8. Initialize $W_{Rj}$with Xavier Initialization
9. **Perform** F=F U R_i& S=S\R_j

b) Support Vector Machine

For the most part, Support Vector Machine (SVM) is viewed as a Classification approach, yet it very well may be connected in the two kinds of arrangement and relapse issues. SVM builds a hyper plane in multi-dimensional space to isolate various classes and it creates an ideal hyper plane in an iterative way, which is utilized to limit a blunder. The center thought of SVM is to locate a most extreme minimal hyper plane that best partitions the dataset into classes.

Algorithm Support Vector Machine

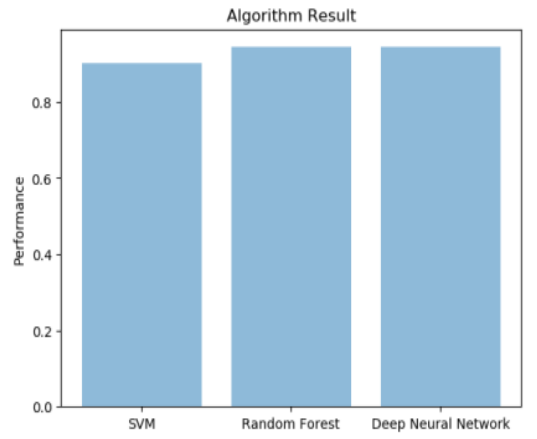
Input:
Sample the Dataset X to classify
Training set S ={(X_1,Y_1)(X_2,Y_2),……(X_n,Y_n)};
The number of nearest neighbors is t
Output:
Decision Y_p€ {-1,1}
Find t
Sample (X_j,Y_j) with minimal values of T(X_j,X) – 2* T (X_j ,X)
Train an Support Vector Machine model on the t selected samples
Classify X using SVM model, get the result Y_p
ReturnY_p.

E.RESULTS WITH ANALYSIS

The experimental setup consists of two components. The first one is collecting data and processing it. The data is collected from Enron email corpus dataset and then the data ispre-processed.The second component being the learning models. Some of the models is simple and some cutting edge. Based on the learning models worked on come up with interesting results.Rapid Table and meta-charts are used to make visual representations. The following analysis shows the different classifiers performance, which achieves more than 80% accuracy than the proposed algorithm

```
plt.ylabel('Performance')
plt.title('Algorithm Result')

plt.show()
```



IV. CONCLUSION

The company's stock price prediction uses computation methods and data mining techniques for exploring stock performance. So as to recognize the huge changes of stock cost, or distinguish the beginning time or hierarchical emergency, the organization can watch the worker's correspondence system and screen its strength.

Enron email corpus dataset is used for analyzing stock price and it is preprocessed with Natural Language Processing. Classification approaches used for analyzing accuracy level of stock price. The comparison of results is visualized with a simple graph format. According to our experimental result, the real stock price can be predicted with the average accuracy more than 80% through our proposed algorithms.In future work, we intend to foresee organization's stock cost in open remarks like those on long range informal communication destinations (for example Twitter and Face book).A future examination will consider demonstrating the correspondence dimension of none, frail, and hearty with the usage of fluffy sets. At last, we will in general infer that it's conceivable to utilize calculation ways like calculations and information mining procedures to investigate a partnership's correspondence examples and use such examples to foresee a company's structure execution like stock execution.

REFERENCES

1. B.Li, K.C.C.Chan, C.Ou,"Public sentiment analysis in Twitterdata for prediction of a company's stock price movements," in Proc.IEEE Int. Conf. e-Bus. Eng. (ICEBE), Nov.2014, pp. 232–239.
2. C. Hargreaves and Y. Hao, "Prediction of stock performance usinganalytical techniques," J. Emerg. Technol. Web Intell., vol. 5,no. 2,pp.136-142,2013.
3. A. A. Sawant and P. M. Chawan, "Study of Data Mining TechniquesUsed for Financial Data Analysis," Int. J. Eng. Sci. Innov. Technol., vol. 2, no. 3, pp.503-509, 2013.
4. M. Gahirwal and M. Vijayalakshmi, "Inter time series sales forecasting," Int. J. Adv. Stud. Comput., Sci. Eng., vol. 2, no. 1, pp. 55–66,2013.
5. B. B. Nair, M. Patturajan, V. P. Mohandas, and R. R. Sreenivasan, "Predicting the BSE Sensex: Performance comparison of adaptive linearelement, feed forward and time delay neural networks," inProc. Int.Conf. Power, Signals, Controls Comput., Trissur, India,jan .2012, pp. 1–5.
6. S. C. Nicholis and D. J. T. Sumpter, "A dynamical approach tostock market fluctuations," Int. J. Bifurcation Chaos, vol. 21, no. 12, pp. 3557–3564, 2011.
7. G. Miller, "Social scientists wade into the tweet stream,"Science,vol. 333, no. 6051, pp. 1814–1815,2011.
8. B. B. Nair, S. G. Sai, A. N. Naveen, A. Lakshmi, G. S. Venkatesh,and V. P. Mohandas, "A GA-artificial neural network hybrid system forfinancial time series forecasting," in Computer and Information Science(Lecture Notes in Computer Science-Communications), vol. 147, V. Das,G. Thomas, and F. Gaol, Eds. Nagpur, India: Springer-Verlag, 2011,pp. 499–506.
9. R. Dash, R. L. Paramguru, and R. Dash, "Comparative analysis ofsupervised and unsupervised discretization techniques," Int. J. Adv. Sci.Technol., vol. 2, no. 3, pp. 29–37, 2011.
10. E. Vamsidhar, K. V. S. R. P. Varma, P. S. Rao, and R. Satapati,"Prediction of rainfall using backpropagation neural network model,"Int. J. Comput. Sci. Eng., vol. 2, no. 4, pp. 1119–1121, 2010.
11. H. Kang, C. Plaisant, T. Elsayed, and D. W. Oard,"Making sense ofarchived e-mail: Exploring the Enron collection with NetLens," J. Amer.Soc. Inf. Sci. Technol., vol. 61, no. 4, pp. 723–744,2010.
12. E. Gilbert and K. Karahalios, "Widespread worry and the stock market,"inProc. 4th Int. AAAI Conf. Weblogs Social Media (ICWSM), 2010, pp. 59-65.
13. Sugumar, R. Rengarajan, A. Jayakumar, C., "A technique to stock market prediction using fuzzy clustering and artificial neural networks", Vol. 33, Issue 5, pp. 992-1024, 2014.