PREDICTION OF STOCK WORTH MOVEMENT USING MACHINE LEARNING

K Geetha[1]
Assistant Professor, School of Computing, SRM Institute of Science and Technology, Chennai, India

B Sreenivas Reddy[2]
Department of Computer Science and Engineering, SRM Institute of Science and Technology Chennai, India

Arun Kumar[3]
Department of Computer Science and Engineering, SRM Institute of Science and Technology Chennai, India

---

ABSTRACT: The aim of this paper is to investigate various strategies for predicting stock worth movement using social media sentiment analysis and data mining. In this article, we will present an effective approach for predicting stock movement with greater accuracy. Social media is an influential forum for people's thoughts and feelings; it's a massive, ever-growing repository of texts ranging from mundane observations to in-depth debates. The aim of sentiment analysis is to extract emotions and feelings from text, and this paper contributes to that area. One of the most basic goals is to categorize text as representing positive or negative sentiment. Sentiment classifiers have been developed for social media text, including product reviews, blog posts, and even tweets. With the the complexity of text sources and subjects, it's time to rethink traditional sentiment extraction methods, and even redefine and enrich the sentiment concept. Then, unlike previous sentiment analysis studies, we construct topical databases within each stream to explore sentiment expression and polarity classification within and across different social media streams. To forecast the demand more accurately, various data mining techniques are used, as well as various hybrid approaches. We conclude that stock prediction is a difficult process that requires consideration of a number of variables in order to forecast the market more correctly and efficiently.

Keywords: Stock Worth Movement Prediction, RF, SVM, LR, NLP, Sentiment Analysis and Machine Learning.

I. INTRODUCTION

Stock market forecasting has long piqued the interest of academics. Despite various scientific attempts, no mechanism for correctly predicting stock market change has been found. The challenges of modeling demand conditions add to the complexity of forecasting. There have been few minor successes despite the lack of reliable prediction techniques.

Fundamental and technical approaches to stock market analysis are two fundamental investing philosophies. Stock Market price fluctuations are thought to be derived from a security's relative data in fundamental research. Fundamentalists predict the future using numerical data such as profits, averages, and management efficacy. Market timing is considered crucial in technical research. Price and volume patterns are identified by technicians using maps and modelling techniques. In order to forecast potential results, these latter individuals depend on historical evidence. Textual data is one field of Stock Market prediction that has had little results. Information from financial reports or breaking news headlines may have a significant impact on a security's share price. The majority of current financial text mining literature depends on a predefined collection of keywords and machine learning techniques. Keywords are usually assigned weights in proportion to the movement of a stock price in these systems. This types of studies have shown a definite, but limited, ability to predict share price course.

II PREVIOUS WORKS

In [13], they used data mining technology to uncover latent trends in historical data that could be used to forecast investment decisions. Stock market forecasting is a difficult job in financial time series forecasting. The stock index is evaluated using five methods: typical price (TP), Bollinger bands, Relative strength index (RSI), CMI, and MA. Using Bollinger Bands instead of MA, RSI, and CMI, the author was able to get a profitable signal of 84.24 percent in this article.

Artificial neural networks were used by Jing Tao Yao et al in [14] for classification, estimation, and recognition. It takes an artist to train a neural network. Trading with neural network outputs, also known as trading technique, is an art form. In this paper, the authors explore a seven-step neural network prediction model construction approach. Skills in pre- and post-data processing/analysis, data sampling, and training critiquing. This article will also include pre and post data processing/analysis capabilities, data sampling, preparation requirements, and model suggestion.

In [15], Tiffany Hui-Kuang et al used neural networks to handle nonlinear relationships and also implemented a new fuzzy time series model to improve forecasting. The Taiwan stock index is forecasted using the fuzzy relationship. In the fuzzy time series neural network model, in-sample observations are used for preparation and out-of-sample observations are used for forecasting. The disadvantage of using all degrees of membership for training and forecasting can have an impact on the neural network's efficiency. Take the gap in observations to prevent this. This narrows the scope of the debate world.

Hidden Markov Models (HMM) were used by Md. Rafiul Hassan et al in [16] to predict stock prices for interrelated markets. Because of its demonstrated suitability for modeling complex systems, HMM was used for pattern recognition and classification problems. The author summarized the HMM's benefit as having a solid mathematical base. It can process new data reliably and efficiently in order to build and test related trends. The author decides to create a hybrid scheme that combines AI paradigms and HMM to increase the precision and reliability of stock market forecasting.

III. PROPOSED WORK

The act of attempting to forecast the potential valuation of a stock using social media is known as stock market prediction. People's opinions and emotions can be expressed freely on social media. Social network research is inextricably linked to sentiment analysis. This is how thoughts and feelings are extracted from text. For analyzing social network content and improving average accuracy, data mining methodologies such as NLP, Random forest, and Neural network are used. Fig.1 illustrates the work flow for stock worth prediction follow as,

i. The daily email based collected data is used for prediction of stock worth movement in this paper.

ii. Next, natural language processing (NLP) is used for preprocess the dataset.

iii. Then, preprocessed dataset is dividing into 70 % of training and 30 % of testing data.

iv. Finally, Random Forest (RF) machine learning approach is used to prediction of stock worth movement.

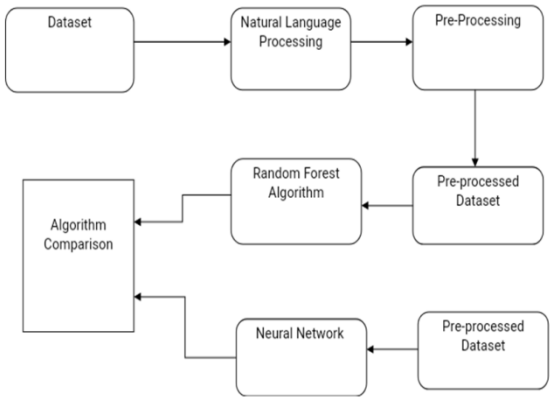

Fig.1 Overall Framework of Proposed System

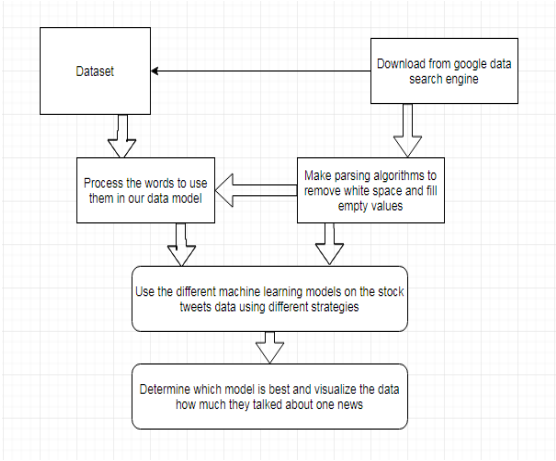

Fig.2 System Architecture of Proposed System

Prediction of Stock Worth Movement implemented as per below modules

1. Data Collection
2. Preprocessing
3. Data Scoring
4. Data Splitting
5. Prediction

1. Data Collection

Our data is primarily comprised of regular stock market emails. We will first attempt to study the NYSE, and then, once perfected, we will apply our model to data from Bangladesh and introduce a framework to assist Bangladeshi investors. We began by looking for tagged data; unfortunately, there were no free sources available. We then created our own site script to scrape chat data from the email API. We run this script for 8 hours a day, and we receive a large volume of chat on a regular basis. Then we parse it and save it in JavaScript Object Notation (JSON) to our file storage.

2. Preprocessing

Our learning algorithm would be hampered by these distracting titles. As a consequence of including these terms, the prejudice seems to rise. If we use these terms, our learning algorithm seems to be looking for quantifiers such as "a," "an," "this," and so on. Auxiliary verbs seem to be of no concern to us as well. So we only need terms that convey the concepts of "positive," "negative," or "neutrality." The following is a list of items we had to do to prepare our data for analysis:

1. Remove all hypertext links from the data in the tweets. "http://" or "https://t.co/3k7Bai5crQ" are omitted from the tweet feed, for example.

2. Lower case all word blocks from the email corpus info. This improves 14 uniformity and allows one to exclude repetitions where they exist.

3. Cleaning up the tweet details by removing white spaces. The emoticons are kept because they have useful information about the tweet.

4. We delete punctuation marks such as commas, full stops, and other punctuation marks.

5. Delete any tag associated with any entity in the email results. Tags that begin with the letter "@" are omitted. We hold hashtags because they can make us interpret the atmosphere better, such as "#Stock #Crash."

6. Use our data email to have a corpus conversation. As a result, tweets with the hashtag "RT" can be removed from the data collection.

3. Data Scoring

Our method for scoring a tweet was straightforward and accurate. The first issue with tweets in CSV files was that they had a large number of loud terms that had little or no meaning in the background. To get terms that concern us, we need to get rid of them. In order to achieve this goal, we began by compiling a list of positive, negative, and neutral terms from the dictionary. Then we graded it based on the tweets' individual positive, negative, and neutral terms as well as our word list. Assume the tweet has n words. Consider the score for positive, negative, and neutral data to be Scorepos, Scoreneg, and Scoreneu, respectively, and the collection of all positive, negative, and neutral terms to be listpos, listneg, and listneu, and the frequency of positive, negative, and neutral words to be frequencypos, frequencies, and frequencyneu, respectively.

4. Data Splitting

In data splitting, divided the seventy percent of it for training set and the rest 30% for test set.

5. Prediction

There are two sections of the experimental setup. The first step is to gather and score results. We rate data gathered from Twitter feeds. The learning models are the second part. Many of the versions are straightforward, while others are cutting-edge. We've come up with some intriguing findings based on the learning models we've previously focused on. To create visual representations, we used rapid Table and meta-charts. We used three learning models on data obtained from a dedicated machine. To build the prediction, three machine learning models are proposed. Three ML models, Support Vector Machine (SVM), Logistic Regression (LR) and proposed method of Random Forest (RF) are used.

A. Support Vector Machine (SVM)

The Support Vector Machine (SVM) is a set of supervised learning methods for classification and regression. The geometric margin is maximized when the analytical classification error is minimized using SVM. Maximum Margin Classifiers (SVM) use the kernel trick to do non-linear classification effectively. An SVM model depicts the examples as points in space, mapped such that the examples of the various classes are separated by a large margin distance. Called training data is given as data points of the sort. The SVM classifier performs classification using an appropriate threshold value after converting the input vectors into a decision value.

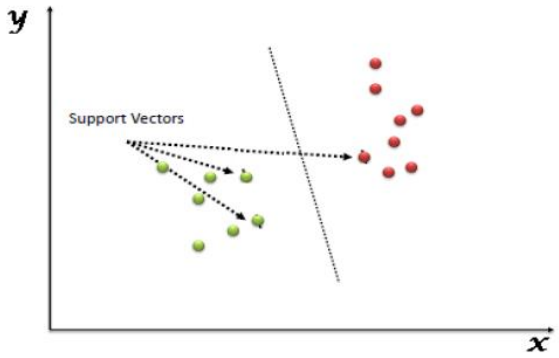

Fig.3 SVM

B. Logistic Regression (LR)

Logistic regression is a method of statistical regression analysis that uses a collection of predictor or independent variables to predict the outcome of a categorical dependent variable. The dependent variable in logistic regression is always binary. The major applications of logistic regression are estimation and estimating the likelihood of success.

C. Random Forest

Random Forest is another common supervised machine learning paradigm. While RF can be used for both classification and regression, classification is its strongest suit. In random forest, decision trees are often used before the output or result is known. As a consequence, "random forest" refers to the combination of many decision trees. A large number of trees can provide a good result in RF. In sorting, the class is determined by a voting system, and in regression, the mean approximation is made for all of the decision tree outputs. Random forest worked well on a wide variety of high-dimensional datasets.

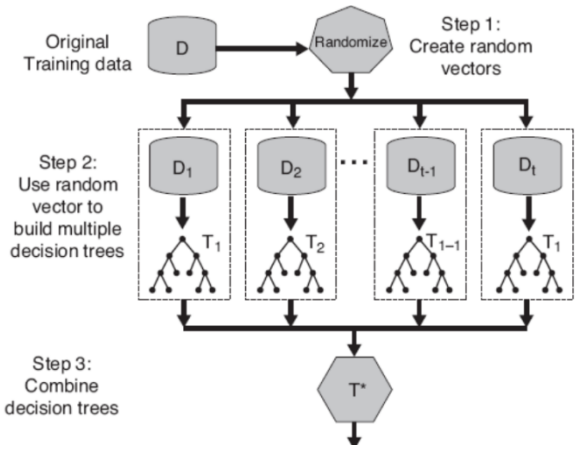

Fig.4. Flow of Random Forest

IV. EXPERIMENTAL RESULTS

In this chapter, we show the prediction results from different ML models. For the construct contrast with three ML models, we used a variety of criteria, one of which is Accuracy.

Table 1: Accuracy analysis of Proposed with different ML models

Algorithms	Accuracy (%)
Support Vector Machine (SVM)	78.72 %
Logistic Regression (LR)	80.12 %
Random Forest (RF)	85.10 %

We can infer from the findings in table 1 that the random forest (RF) model has a higher precision of 85.10 percent than the other two models.

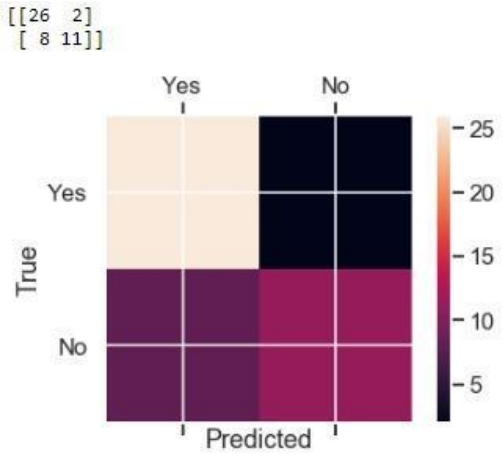

Fig.5 Confusion Matrix for SVM

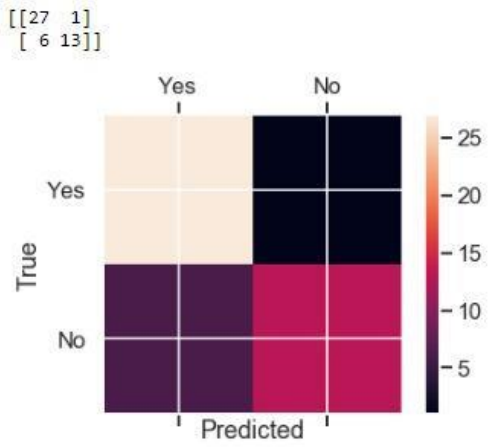

Fig.6 Confusion Matrix for RF

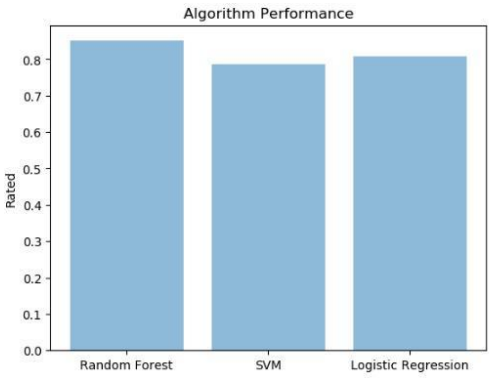

Fig.7 Accuracy Analysis

Fig.7 shows the performance analysis to three machine learning models.

V. CONCLUSION

As a result, we investigated numerous sentiment analysis and data mining approaches for stock market prediction. Opinion mining has become a common research topic due to its utility and demand from the

public. Analysis and summarizing opinionated data is becoming more relevant as the amount of opinionated data grows. This paper proposes three machine learning-based approaches for predicting stock worth movement: support vector machine (SVM), logistic regression (LR), and random forest (RF). The prediction findings showed which method is best suited for predicting stock worth movement by comparing their accuracy.

REFERENCES

[1] Wenping Zhang, Chunping Li, Yunming Ye, Wenjie Li and Eric W.T. Ngai "Dynamic Business Network Analysis for Correlated Stock Price Movement Prediction", Published by the IEEE Computer Society 2015

[2] Chia-Hsuan Yeh and Chun-Yi Yang "Social Networks and Asset Price Dynamics", IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, , JUNE 2015

[3] Li-Xin Wang "Dynamical Models of Stock Prices Based on Technical Trading Rules Part I: The Models", IEEE TRANSACTIONS ON FUZZY SYSTEMS, AUGUST 2015

[4] Xiaodong Li a, Haoran Xie a,⇑, Li Chen b, Jianping Wanga, Xiaotie Deng "News impact on stock price return via sentiment analysis", ScienceDirect 2014

[5] Hong Keel Sul,Alan R. Dennis,Lingyao (Ivy) "Trading on Twitter: The Financial Information Content of Emotion in Social Media", 2014 47th Hawaii International Conference on System Science

[6] Abbasi, H. Chen, A. Salem, Sentiment analysis in multiple languages: feature selection for opinion classification in web forums, ACM Trans. Inform. Syst.(TOIS) 26 (2008) 12.

[7] S. Bao, S. Xu, L. Zhang, R. Yan, Z. Su, D. Han, Y. Yu, Mining social emotions from affective text, IEEE Trans. Knowl. Data Eng. 24 (2012) 1658–1670.

[8] M. Bautin, L. Vijayarenu, S. Skiena, International sentiment analysis for news and blogs, in: Proceedings of the International Conference on Weblogs and Social Media, 2008.

[9] F. Bießmann, J.M. Papaioannou, M. Braun, A. Harth, Canonical trends: detecting trend setters in web data, in: International Conference on Machine Learning, 2012.

[10] E. Cambria, C. Havasi, A. Hussain, Senticnet 2: a semantic and affective resource for opinion mining and sentiment analysis, in: FLAIRS Conference, 2012, pp. 202–207.

[11] E. Cambria, T. Mazzocco, A. Hussain, Application of multi-dimensional scaling and artificial neural networks for biologically inspired opinion mining, Biol. Inspired Cogn. Architec. 4 (2013) 41–53.

[12] E. Cambria, B. Schuller, Y. Xia, C. Havasi, New avenues in opinion mining and sentiment analysis, IEEE Intell. Syst. (2013).

[13] K. Senthamarai Kannan, P. Sailapathi Sekar, M.Mohamed Sathik and P. Arumugam, "Financial stock market forecast using data mining Techniques", 2010, Proceedings of the international multiconference of engineers and computer scientists.

[14] JingTao YAO and Chew Lim TAN , "Guidelines for Financial Prediction with Artificial neural networks".

[15] Tiffany Hui-Kuang yu and Kun-Huang Huarng, "A Neural network-based fuzzy time series model to improve forecasting", *Elsevier*, 2010, pp: 3366-3372.

[16] Md. Rafiul Hassan and Baikunth Nath, "Stock Market forecasting using Hidden Markov Model: A New Approach", Proceeding of the 2005 5th international conference on intelligent Systems Design and Application 0-7695-2286-06/05, IEEE 2005.

[17] Ching-Hsue cheng, Tai-Liang Chen, Liang-Ying Wei, " A hybrid model based on rough set theory and genetic algorithms for stock price forecasting", 2010, pp. 1610-1629.

[18] M.H. Fazel Zarandi, B. Rezaee, I.B. Turksen and E.Neshat, "A type-2 fuzzy rule-based experts system model for stock price analysis", Expert systems with Applications, 2009, pp. 139-154.