Efficacy of News Sentiment for Stock Market Prediction

Sneh Kalra , Jay Shankar Prasad
Department of Computer Science Engineering
MVN University, Palwal, India
snehchhabra@gmail.com, jayshankar.prasad@mvn.edu.in

Abstract- Stock Market trend prediction will always remain a challenging task due to stochastic nature. The enormous amount of data generated by the news, blogs, reviews, financial reports and social media are considered a treasure of knowledge for researchers and investors. The present work focuses to observe fluctuations in stock prices with respect to the relevant news articles of a company. In this paper, a daily prediction model is proposed using historical data and news articles to predict the Indian stock market movements. Classifier Naïve Bayes is used to categorize the news text having negative or positive sentiment. The count of the positive and negative sentiment of news articles for each day and variance of adjacent days close price along with historical data is used for prediction purpose and an accuracy ranging from 65.30 to 91.2 % achieved with various machine learning techniques.

Keywords: News Sentiment, Sentiment analysis, Machine learning, Stock price prediction, social media analytics

I. Introduction

The stock price fluctuation affects an individual's life in financial as well as other segments. For selection of forecasting methods, accuracy is the most essential factor. Forecasting methods believe that publicly available information in the past on the social media, news articles, and financial reports has a strong relationship to the future stock return [20]. There is a need to find the appropriate forecasting method to find the best results.

Positive News encourages individuals to buy and sell stocks respectively [11]. New products and acquisitions, better-quality earnings reports, overall economic and political indicators increase the demands and stock price. In contrast, negative news leads to decrease in demands and stock price and cause individuals to sell stocks. Economic and political uncertainty, poor corporate governance, poor earnings reports, unexpected, unfortunate occurrences will turn to sell demands and a decrease in stock price [11, 18].

This paper reveals the relationship between movements in stock markets and financial news. The news articles about a company and general news of the stock market have a significant impact on the stock movement. Therefore stock related keywords are used for filtering the news to find the genuine impact of news on the company's stock price [17].

In the proposed model, count of sentiments values of news articles, historical stock data, and variance of adjacent days close price are used for future price movement prediction. The initial phase is the analysis of news articles to find its text polarity. The second phase combines the historical data, variance and numeric sentiments values together to predict stock prices.

The contribution of this research lies in the use of existing classification and prediction algorithms to the dataset. The dataset consists of news dataset and stock price dataset.

II. Literature Review

Numerous studies had explored stock market movement prophecy using news and social media analytics.

SA Bogle and WD Potter [1] found social media tweets comments have significant effects on the stock market of Jamaica stock exchange. Generally, the tweets are unstructured and require a pre-processing routine to perform the sentiment analysis. The stock predictions of [1] used machine learning predictors such as neural networks, support vector machine and decision trees on news sentiment data. An accuracy of 87% in the motion prediction and value of correlation coefficient 0.99 for price prediction has been reported which suggests the scope of improvement [1]. Similarly, a robust relationship among stock prices of a company to the emotions or public opinions about the company articulated on twitter by analyzing tweets sentiments has been reported in [2]. Textual representation of Ngram and Word2vec for finding the public sentiments in tweets using logistic regression [2] is another approach. The relation between market sentiment and public sentiment can also be seen in [3] which use twitter data for public mood prediction and used previous day's values for predicting the movements.

Forecast model based on historical stock market prices and sentiment analysis of financial news is reported in [5]. The model achieved accuracy results ranging 72-86.21% by considering historical stock price and multiple types of news related to company and market [4]. The prediction accuracy can be improved by combining historical stock prices with news polarities [5]. Similarly for predicting Indonesian market stock movement based on sentiment values of tweets, margin percentage prediction, stock price prediction and price fluctuation prediction methods has been reported on [6]. It is proved that random forest and naïve bayes classification algorithms of prediction performed well in comparison to the other used classification algorithms [6]. However, for the prediction purpose, the prices of five previous days are very much useful. Bing used a model to analyze hourly stock prices trend and public tweets [7]. Data mining techniques along with NLP techniques to find relationship patterns among numeric stock prices and public sentiment is reported in [7]. The sentiment analysis on news articles to influence the share price is reported in [8] by collecting the news dataset using Bing API. A specialized sentiment dictionary to analyze stock articles is a good approach to handle the prediction strategy as reported in [8]. Qing Li et al. [9] found that firm-specific news articles can enrich the knowledge of investors by implementing a proposed quantitative media-aware trading strategy to see its impact on stock markets. Sentiments lead to emotional fluctuations in investors to make decisions [8, 9]. Article content and firm characteristics also affect trading activities [8, 9]. The daily and monthly prediction model used in [10] was based on supervised machine learning algorithms. For the daily prediction model, sentiments were combined with historical data. For monthly prediction model, two months trends were used to find any similarity between them. Findings proved that one month's trend is least correlated with another month's trend [10].

The related literature suggests that the prediction of stock price depends on several factors and gross impact lies on the various modalities. Hence, each and every aspect can be given weightage while designing any machine learning model for stock market trend prediction.

III. Proposed Method

To enhance the prediction accuracy of stock market movement, the proposed system combines the sentiment values of news, historical data, and variance of adjacent days close price. The proposed model helps to make the decision easier for the investors to avoid financial loss and risks while investing in the stock market. The model forecast the price movement on a day by taking into account all the available news and numeric historical data. Supervised machine learning techniques are used to train the available data. News sentiments are extracted and combined with numeric historical price to build the prediction model. This study performs text analysis on news data to find text polarity.

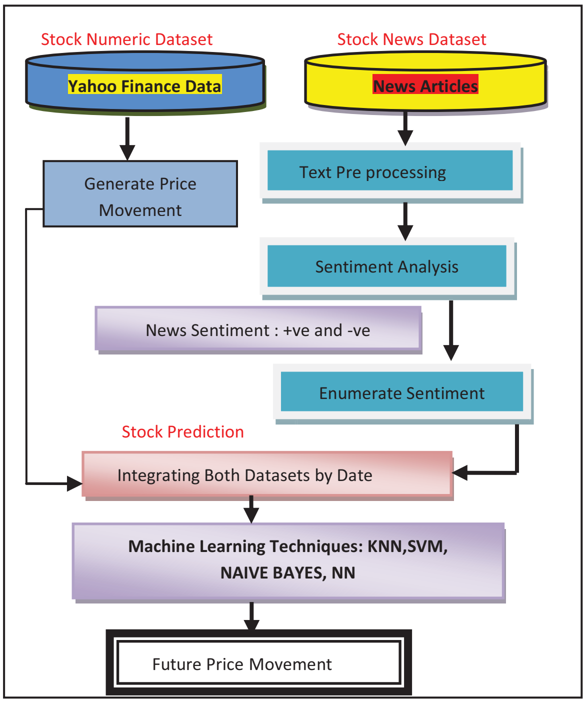

Fig.1. Proposed Prediction Model for the Stock Market

Furthermore, Stock historical prices open, low, high and variance of adjacent days close price are used for future price movement prediction. Fig. 1 represents the proposed prediction model for the stock market and self explanatory.

A. *Stock Numeric Dataset*

The numeric dataset contains the attributes low, open, high, adjacent close price which is taken from Yahoo Finance and variance of adjacent days close price is calculated.

For historical data prices, price movement for a day is generated using the adjacent close price of a stock and difference between adjacent days close price is calculated by subtracting previous day close price from today's close price shown in Algorithm [10]. If obtained value is non -negative number then trend is up otherwise down. The open, high, low attributes are used in their numeric form. For the considered time period, the news data was available for all days, stock data was not available when the market is closed including weekends and holidays [3]. To find the missing value a variable named α is used. α is calculated as median of all available values and we substituted the missing values by calculated value of α.

The Algorithm to Generate Price Movement

```
Input   :   Adjacent Close Price Value (PV)
Output  :   Price movement  for each day(m)
    Diff← (Today PV- Previous day PV)+ α
          concatenate  0 to Diff
          initialise  j=0;
    for   i in Diff do
                  if i>0 then
                            m[j]=up;
                    else
                  m[j]=down;
   end for
         j=j+1;
  end;
// α = Median of all available data
```

B. Stock News Dataset

For news, data is collected from online financial websites on daily basis, using XML mappings such as moneycontrol, livemint, financialexpress, businesstoday and ndtv. The collected news data is from different sections of the newspaper such as banking and finance, business news, buzzing news, stock indices, economy, market edge, economy etc. The attributes collected in news data are Title, Link, Comments, Publication date, Creator and Description. For the sentiment purpose, we have considered the attribute Title.

A. Sentiment Analysis

To find the sentiment of news, the news data is analyzed and categorized into positive and negative sentiments. To attain the required results pre-processing of data is done on news dataset and naive Bayes algorithm [6] is used for news classification. The details of performed steps are as follows

- Text Pre Processing : To improve the accuracy of sentiment analyzer following pre-processing steps are performed:

Data Cleansing: The news articles collected from various financial sources include the data that was unrelated to the stock market. Specific company related news articles are kept in the dataset. To find stock related news[15], the numbers of words such as share market, rupee fall, dollar, sensex considered for filtering the remaining news from stock news dataset.

Tokenization: Tokens are generated for each news article by splitting it into number of words.

Transformation: Transformation is done by converting all the words in news articles in a document into lower case.

Stopword Removal: Words that do not express any meaning such as "a , the, an , is " etc are removed from splitted news data .

Term Frequency - Inverse Document Frequency: Term frequency-inverse document frequency is a procedure that considers a term's frequency (TF) and its inverse document frequency (IDF) to find the importance of words in a document. Equation (1) is used to calculate TF- IDF [5] where N is the total number of documents and DF_t is the number of documents containing the term t and $TF_{t,d}$ is the number of occurrences of t in document d .

$$TF - IDF = TF_{t},_{d} * \log(N/DF_{t}) \qquad (1)$$

- Naive Bayes Classifier: The Classifier is trained with the training data set and implemented on testing data to know the sentiments of news dataset[6,5]. For the sentiment purpose, the given data is divided into training and testing Data.

Training Data: Training dataset assists the machine to learn how to perform when it is given a set of inputs. Approximately, 70 % of available data is used as training data and sentiment is assigned to each news article. The assigned sentiment is having positive or negative polarity.

Testing Dataset: Testing dataset helped the classification algorithm to forecast the sentiments of remaining 30% data. Naive Bayes classification algorithm is applied to forecast the sentiments of the testing dataset.

B. Enumerate Sentiment Values

For a single day, there might be more than one or two genuine stock related news for a company. The calculated values of positive and negative sentiments for each day are converted to numeric form. Instead of considering the news text polarity as positive and negative we have enumerated positive sentiment values and negative sentiment values for each day.

C. Prediction Model

The proposed prediction model integrates numeric news sentiments values, variance and numerical stock prices to observe the influence of released news, variance and historical data on stock movements. The two components of the prediction model are described as follows.

- Integration of Datasets by Date: The up or down price movement for a day and numeric sentiment value combined together on daily basis resulting in open, low, high, variance, price movement, positive sentiment count and negative sentiment count . It works as input for the prediction model.

- Stock Prediction (Machine Learning Predictors): For predicting the stock price movement based on the integrated dataset, Machine learning techniques are applied for classifying the dataset. KNN[5], Neural Network (NN)[20], Support Vector Machine[13](SVM) and Naive Bayes are applied to find the predicted class. These algorithms are compared depending on their performance.

IV. RESULT AND PERFORMANCE ANALYSIS

The implementation phase works in two parts. The first part produces the results of news sentiment that gives polarity of news data as positive or negative. The second part generates the result of the forecast model that takes count of positive and negative news, variance with historical data as input. For experiment purpose, datasets of Bank of Baroda (BOB), Punjab National Bank(PNB), HDFC Bank and ICICI Bank for August 2018 and September 2018 traded in NSE have been collected. News data for HDFC Bank, ICICI Bank, PNB and BOB has almost 309 rows, 300 rows, 302 rows, and 245 rows respectively. For historical data, all the days were considered except Sunday. Rapidminer tool is used for the implementation purpose.

A. Sentiment Analyser Result

The sentiment analysis using Naïve Bayes is classifying the text as positive and negative. Fig. 2 represents the sentiment result for HDFC Bank with TF-IDF.

B. Prediction Model Result

To predict the future price movement, an integrated dataset is divided into training and testing datasets. The prediction model shows that numeric sentiment values and variance with open, high and low value improving the prediction accuracy up to 91.2 % using KNN algorithm. The results also prove that variance attribute has a high influence on stock price movement.

Four machine learning techniques named KNN, SVM, Naive Bayes, and Neural Network are used for the prediction purpose. Table I, table II, table III and table IV represents the prediction accuracy for ICICI Bank, HDFC Bank, PNB and BOB using above mentioned machine learning algorithms.

1. Prediction accuracy achieved for KNN varies between 75% to 91.2%.
2. Prediction accuracy achieved for SVM varies 65.30% to 83.80%.
3. Prediction accuracy achieved for Naive Bayes varies between 74.70% to 85%.
4. Prediction accuracy achieved for Neural network varies between 73.80% to 88.70%.

Numeric count of sentiments values and variance make a good prediction difference in comparison to previous studies which has achieved accuracy between 65% to 86.12 %. Fig. 3 represents the performance comparison of four classifiers. Our proposed model is a helpful approach to improve the prediction accuracy using numeric count of sentiments and variance with historical data. The proposed model is compared with result of previous studies and our model is performing the most excellent among all studies. Table V represents the comparison of result with state of art methods.

V. CONCLUSION AND FUTURE WORK

The implemented model examines the effect of analyzing diverse types of stock related news with numeric historical data on the stock market. An effort is made to build a useful model to predict the future price movement. The future price movement accuracy is improved by the considering numeric sentiment value and numeric historical data. The highest prediction accuracy achieved with KNN is 91.2%. The result of the proposed method is well-matched with previous researches that state that there is a strong correlation between stock related news and change in stock price. On the considered dataset KNN is performed best in comparison to other applied algorithms. The future work may consider social media data , reviews ,blogs for a long time period that influence the stock market and by considering high number of news data instances.

S.No	News Text	Sentiment	Rupee	Dollar	RBI	Raise	Plunge	drop
1	HDFC Bank raises Rs 15,151 cr from domestic, overseas mkt	Positive	0	0	0	2.322	0	0
2	Stock Market LIVE: Sensex flat, rupee plunge past 70.50 against US dollar	Negative	0.257	0.344	0	0	1.367	0
3	Closing bell: Sensex dive 188 points dragged by banking stocks, Nifty settles at 11,385	Positive	0	0	0	0	0	0
4	Stock Market LIVE: Sensex edges lower, rupee falls beyond 70.80 against US dollar	Negative	0.257	0.344	0	0	0	0
5	Street signs: HDFC Bank ADR premium drops	Negative	0	0	0	0	0	2.021
6	HDFC Bank tells Sebi it was unaware of conflict of interest	Negative	0	0	0	0	0	0

Fig. 2. Sentiment Result for HDFC Bank with TF-IDF

Table I. Prediction Results for ICICI bank

Machine Learning Algorithms	**ICICI Bank**			
	% Accuracy	***Precision***	***Recall***	***F Measure***
KNN	78.7	0.788	0.787	0.787
SVM	81.2	0.817	0.812	0.812
Naive Bayes	80	0.801	0.806	0.800
Neural Network	80	0.801	0.800	0.800

Table II. Prediction Results for HDFC Bank

Machine Learning Algorithms	**HDFC Bank**			
	% Accuracy	***Precision***	***Recall***	***F Measure***
KNN	75.0	0.754	0.750	0.744
SVM	78.7	0.785	0.787	0.783
Naive Bayes	80	0.801	0.806	0.800
Neural Network	73.8	0.736	0.738	0.736

Table III. Prediction Results for Punjab National bank

Machine Learning Algorithms	**Punjab National Bank**			
	% Accuracy	***Precision***	***Recall***	***F Measure***
KNN	85.3	0.853	0.853	0.852
SVM	65.3	0.587	0.685	0.653
Naive Bayes	74.7	0.747	0.748	0.747
Neural Network	80.0	0.800	0.800	0.796

Table IV. Prediction Results for Bank of Baroda

Machine Learning Algorithms	**Bank of Baroda**			
	% Accuracy	***Precision***	***Recall***	***F Measure***
KNN	91.2	0.913	0.912	0.912
SVM	83.8	0.843	0.838	0.837
Naive Bayes	85	0.850	0.854	0.850
Neural Network	88.7	0.890	0.887	0.887

Table V. Comparison of Result with State of Art Methods

Previous Studies	Accuracy
Vaanchitha Kalyanaraman et al model [8]	81.81%
Ayman E. Khedr et al model [5]	86.21%
Arman Khadjeh et model [15]	65.20 %
Ventaka Sasank Pagolu et al model [2]	69.01 %
Proposed model	91.20%

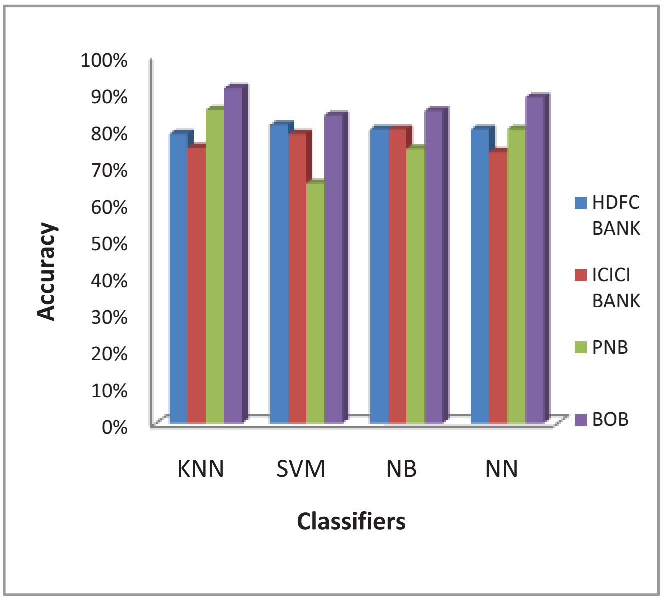

Fig. 3. Performance Comparison of Four Classifiers

REFERENCES

[1] S. A. Bogle , W.D. Potter, "SentAMaL - A Sentiment Analysis Machine Learning Stock Predictive Model," in Proc. of Int. Conf. of Data Mining and Knowledge Engineering , 2015.

[2] V. S. Pagolu, K.N. Reddy, G. Panda , B. Majhi , "Sentiment Analysis of Twitter Data for Predicting Stock Market Movements," Int. conf. on Signal Processing, Communication, Power and Embedded System(SCOPUS) 3-5 Oct. 2016.

[3] A. Mittal , A. Goel , "Stock Prediction Using Twitter Sentiment Analysis," Standford University, 2012.

[4] P. S. Michael Rechenthin ,W. N. Street, "Stock Chatter: Using Stock Sentiment to Predict Price Direction, "Algorithmic Finance, vol. 2, no. 3-4, pp. 169-196, 2013.

[5] A. E. Khedr , S.E. Salama , and N. Yaseen , "Predicting Stock Market Behavior using Data Mining Technique and News Sentiment Analysis," Int. J. Intelligent Systems and Applications, pp. 22-30 , 2017.

[6] B. D. Trisedya, Y. E. Cakra, "Stock Price Prediction using Linear Regression based on Sentiment Analysis", Int. Conf. Adv. Computer. Sci. Inf. Syst. , pp. 147–154, 2015.

[7] L. I. Bing , K. C. C. Chan , C. Ou, "Public Sentiment Analysis in Twitter Data for Prediction of a Company's Stock Price Movements," IEEE 11th Int. Conf. E-bus. Eng. pp. 232-239, 5-7 Nov. 2014.

[8] V. Kalyanaraman , S. Kazi , Rohan, Tondulkar , S.Oswal , " Sentiment Analysis on News Articles for Stocks," 8th Asia Modelling Symposium, 23-25 Sept. 2014.

[9] Q. Li , T. Wanga, P. Li , L. liu, Q. Gonga, and Y. Chenb, " The effect of news and public mood on stock movements ," Information Science , Vol. 278, pp. 826-840, Sep 10, 2014.

[10] A. Nayak, M. M. Manohara Pai, R.M. Pai "Prediction models for Indian stock market," Procedia Computer science, Vol. 89, pp. 441-449, 2016.

[11] D. K. Kirange , Dr. Ratnadeep , and R. Deshmukh , "Sentiment Analysis of News Headlines for Stock Price Prediction," COMPUSOFT, Int. J. advanced computer technology, Vol. V, Issue-III ,March 2016.

[12] F.J. Garcia-Lopez ,I. Batyrshin, and A. Gelbukh, "Analysis of relationships between tweets and stock market trends," Journal of Intelligent and Fuzzy Systems , May 2018.

[13] A. Porshnev , I. Redkin , and A. Shevchenko, "Machine Learning in Prediction of Stock Market Indicators Based on Historical Data and Data from Twitter Sentiment Analysis," IEEE 13th International Conference on Data Mining Workshops, 7-10 Dec. 2013 .

[14] Y. Kim, S. R. Jeong, and I. Ghani, "Text Opinion Mining to Analyze News for Stock Market Prediction," Int. J. Adv. Soft Comput. Its Appl., vol. 6, no. 1, pp. 1–13, 2014.

[15] A. Khadjeh, Nassirtoussi, S.R. Aghabozorgi ,Y.W. Teh, and D.C. Ling Ngo, "Text mining for market prediction: A systematic review," Expert System Appl., vol. 41, pp. 7653-7670, 2014.

[16] I. Bernardo, R. Henriques, V. lobo , "Social Market: Stock Market and Twitter Correlation," Int. Conf. on Intelligent Decision Technologies, 26 May 2017.

[17] Q. A. Al-Radaideh, A.A. Assaf , and E. Alnagi, "Predicting Stock Prices Using Data Mining Techniques," Int. Arab Conf. on Information Technology, 2013.

[18] R. Ahuja, H. Rastogi, A. Choudhuri and B. Garg, "Stock Market Forecast Using Sentiment Analysis," 2nd Int. Conf. on Computing for Sustainable Global Development , pp. 1008-1010, 11-13 Mar. 2015.

[19] V. Ingle, S. Deshmukh, "Hidden Markov Model Implementation for Prediction of Stock Prices with TF-IDF features ," in Proc. of the Int. Conf. on Advances in Information Communication Technology & Computing, 12-13 Aug. 2016.

[20] L. Ertuna, "Stock Market Prediction Using Neural Network Time Series Forecasting," May 2016.