Short Term Stock Movements with Big Data and Market Sentiments Analytics

Sneh kalra, Sachin Gupta, Jay Shankar Prasad

Abstract*: Machine Learning Techniques and Big Data analytics are two central points of data science. Big Data is important for organizations to gain insights into it and machine learning techniques are one of the substantial assets for analyzing a massive amount of data. In this paper, a framework has been proposed to improve the short term stock trend prediction accuracy using Logistic Regression model by means of qualitative and quantitative data. This paper makes a comprehensive survey of stock market trend prediction with the accumulation of various data sources by applying machine learning techniques and by using big data analytics approach. The model has been implemented in Big data Framework with Hadoop and Apache Spark. For qualitative data Tweets sentiments and news sentiments has been taken in to account and for quantitative data Google trends and historical data are considered. The proposed system has enhanced the prediction accuracy about 3-4 % in comparison to existing models by supplying Google trend as input data in addition to market sentiments and historical data. The implemented model can help the investors to take short term decisions to make money in the security market and the survey would help in finding the most effective resources which overly influence the stock prices.

***Index Terms*: Google trends, Logistic Regression, News sentiments, Sentiment Analysis, Tweet sentiments**

I. INTRODUCTION

The stock exchange is fully packed with unreliability and influenced by numerous factors [13]. Share market is one of the leading unpredictable places of high interest within the finance and business [4]. There are not any vital ways exist to predict the share price. The foremost necessary juncture of this investigation lies in making prognosis(prediction) concerning real stock market prices with disparate information sources[20] using machine learning techniques [22,23]. A few years back, in the absence of a huge amount of social media data, researcher or investors used to make predictions by means of exploring time series data [5]. The share market data does not repeat periodically, the characteristics of data are not consistent and are insignificant to time series data. So to collect features only from past data is a tough job. Time-series analysis does not work up to the mark for the stock market prediction. Certainly, the results of such attempts were not greatly favorable which imposed the authors to proceed towards machine learning algorithms [22, 23]. So there is a noteworthy necessity for tools that can contribute to outstanding prediction quality from social media data and published news records.

If online news or tweets for a specific firm or company is positive, capitalist get the firm's stock to obtain profit else sell the equivalent stock [9]. It's forever ascertained that stock exchange values are always reaching the utmost low or maximum top values if financial news is broadcasting headlines [9].In last twenty years, with the expansion of storing and trailing systems [13], an enormous quantity of past data is out there for examination hence machine learning techniques elected as the main weapon for innovative jobs. Previously, several different machine learning techniques are applied with varying level of achievement. Though, stock forecasting remains strictly limited due to its volatile and unpredictable nature.

As the stock price is affected by many factors including news, political events, natural calamities and social media events [18]. The data generated by these sources are very huge and needs to get analyzed to see how it can affect the stock market. As this data is unstructured so along with analysis, this data needs special handling for its storage and processing [15]. For the purpose, Researchers move towards big data approach for the prediction task. In this paper, a new framework has been proposed for stock prediction with social media data and news sentiments by implementing Logistic Regression in Big Data environment. In addition to that, the relation between the Google trend (what an individual thinks while searching over the internet) and the stock market trend is also investigated.

II. LITERATURE REVIEW

The main focus of the literature study is to characterize the connection between input data choice and stock market prediction accuracy by addressing specific studies supported financial news sentiments and twitter sentiments using machine learning classifiers [21, 22, 23] and few studies based on historical data, tweets sentiments and news sentiments by means of big data analytics.

In this review study, for every publication, four categories are specified. Those categories are Authors, the Methodology used, Forecasted Index, Input variables, and Results. In the "Author & Year" category, name of the authors that have conducted the research and year of the research are listed. The second category is "Methodology" describes the techniques and algorithms used in the specific study. The other category named "Forecasted index" describes the market situations like developed markets, rising markets and, frontier markets. Input datasets/variables or fact-finding variables are main parameters for a prediction model because the prophetical power of the model is largely dependent on the inputs used hence the

Revised Manuscript Received on July 09, 2019.

Sneh Kalra, Computer Science, MVN University, Palwal, India.

Dr. Sachin Gupta, Computer Science, MVN University, Palwal, India.

Dr. Jay Shankar Prasad, Computer Science, Krishna Engineering College, Ghaziabad, India.

third category. The last essential class describes the result of studies in terms of accuracy of prediction, error rate or performance measure achieved by proposed prediction models is the "Result" category. All of the reviewed papers are summarized within the table-1.1and table-1.2 based on the selection of input variables that have been considered for prediction.

A. Sentiments and Machine Learning Techniques

Analysis in behavioral financial system evidently specifies that news sentiment is considerably associated with stock price movements. In addition, researchers try to take advantage of the information available publicly in social media and particularly, different researches showed an association between twitter messages and the stock exchange index. This section covers a few studies making use of news and tweet sentiments with machine learning algorithms [19].

The authors in [1] worked with historical data and news articles of Wipro and Infosys that has been collected for 10 years. Their study proved that relationship exists among the stock market and positive polarity of news articles by finding the correlation. For experimentation, they have trained support vector machine (SVM) with word count dictionary and linguistic inquiry. Naive Bayes, k Nearest Neighbour (knn) classifiers were employed for emotion classification of financial news. Findings reported an average accuracy of 90.51 %, 67.13 %, 63.91 % for support vector machine, k nearest neighbour and naive bayes classifiers [21] respectively.

The authors in [2] predicted the direction of returns using news articles generated from social media. News articles have collected for stocks traded under NSE. Further, the authors found the causation relation between news articles and stock prices. They conducted the experiment for finding sentiment and resulted in an accuracy of 53.2%, 59.5% with Normal Equation and Gradient Descent. Then they compared prediction results obtained with decent with Normal Equation and found that Gradient Descent is performing well in comparison to the Normal Equation.

The authors in [3] used public mood states extracted from microblog feeds with historical data to predict Chinese stock market movements. Granger Causality test was conducted to confirm the relation between price and mood states. Algorithms, probabilistic neural network (PNN) and SVM implemented for prediction and results revealed that SVM performed better than PNN. Furthermore, findings also show that by providing two dimensions to input data resulted in the highest accuracy of 71.42 % in comparison to accuracy that is achieved by using only historical variables as input data. The results were also compared with ROSTEA scores and found to be more accurate than ROSTEA results.

The authors in [4] compared the prediction accuracy of different machine learning algorithms with k fold methods and observed that SVM and Random Forest performed better than Naive Bayes. For the experimentation, sentiments values of news articles with historical data used as input to implemented algorithms. They also reported that the proposed model enhanced the prediction accuracy by 30 %.

The authors in [5] used twitter sentiment values in combination with commodity prices, analyst's recommendations which worked as input to the neural network model. Usage of commodity prices, the analyst's recommendations resulted in a 9% increase in prediction accuracy. Furthermore, they conducted Granger Causality test to observe the association between stock prices and twitter sentiments, correlation test to observe the association among real stock prices and product prices and t-test to observe the relation among recommendations and stock prices.

The authors in [6] trained SVM with linear kernel for predicting the stock prices. They made use of social media extracted sentiments and special parameters extracted from the yahoo finance message board. Their proposed model was able to discover existing topics and their associated sentiments automatically. They reported a 2.07% enhanced accuracy in comparison to the accuracy achieved by using only historical prices.

The authors in [7] tested the potential profitability of Raven Pack Dow Jones (RPDJ) News analytics for Google stock trends by developing an Artificial Neural Network prototype. Using the unique news collection source, the proposed model resulted in prediction rate of 54.47 % and proved that ANN model performed superior in comparison to a random walk directional prediction system, which states that the stock value has a 50-50 % possibility of closing lower or higher in comparison to the stock opening price.

The authors in [8] collected data from micro-blogger site, chat rooms, and web forums to investigate the relation among sentiments and stock price fluctuations in China. They conducted correlation and causality tests among multiple post sentiment scores and stock price return and trading volume. Their causality results show chat room post sentiment score and stock price trends are closely related to each other and also proved the existence of causality between them. Further, they used chat room sentiments for experimentation using NBSVM, LSTM, Ensemble methods and at the end of 7 months achieved a return of 19.54%.

The authors in [9] worked with online news data and stock market parameters named as close, high, low, open, etc. To predict the next day stock value for some of the companies, the Hidden Markov Model implemented using TF-IDF features. Their result shows an error rate of 0.2 to 3.9 % for the closing price for a week.

The authors in [10] extracted Apple product tweets from Stockwits and conducted a sentiment analysis process on them. Along with tweets sentiments score, the author collected historical data and served both data as input variables to SVM for predicting next day stock movement. Her findings resulted in a positive association among market data and people opinion and her model resulted in 76.65% accuracy for the prediction task. Table I shows detailed summary of studies based on sentiments with machine learning techniques.

Table I Summary of studies based on Sentiments with machine learning techniques.

S.No	Author Name and Year	Methodology Used	Forecasted Index and Input Datasets/Variables	Results
1	D. K. Kirange et al. (2016)	• Emotion classified using Naive Bayes, knn, and SVM. • Feature matrix constructed using LIWC.	• Infosys and Wipro • 10-year news dataset, Historical data	• Emotion classifier achieved accuracy of 90.51 %, 67.13 %, 63.91 % for Infosys datasets and 75.45 %, 47.98 %, 72.64 % for Wipro datasets using SVM, KNN and Naive bayes. • Achieved maximum correlation accuracy of 94.44% between positive % of news and stock prices. • Used svm to established contrast among positive sentiment curve and stock price trends.
2	V. Kalyanaraman et al. (2014)	• Sentiment analysis was done using feature extraction & specialized sentiment dictionary. • Linear Regression techniques named Gradient Descent and Normal Equations used for prediction.	• 11 companies from NSE. •Around 1000 news articles, historical data.	• Achieved 54.54 % and 81.81 % prediction accuracy. • Predicted the sentiment of a news article with 53.2% and 59.5% accuracy.
3	Danfeng Yan et al. (2016)	• Proposed Chinese profile of mood state for microblog feeds to analyze their sentiment. • Performed Granger Causality test to confirm the relation between C-POMS analysis and price series. • Prediction made using SVM and pnn (Probabilistic neural network)	• Chinese stock market. • Public moods feed (644929 feeds), Historical data.	• Results indicated the improvement in prediction accuracy up to 66.67 % using C-POMS as the input. • The model achieved 20 % more accuracy with two input dimensions in comparison to model using only historical data as input.
4	Kalyani Joshi et al. (2016)	• Created polarity words dictionary based on McDonald's research. • Classifiers SVM, RF, Naive Bayes. • Evaluation using 5-fold, 10-fold, 15-fold validation methods for 70%, 80% data split and for New testing data.	• Apple Company. • 3 years of News articles, Historical data.	• Classifiers SVM, RF, Naive Bayes achieved 90%, 86% and 83 % for unknown data and accuracy ranging between 81.52% -94.44% for 5-fold, 10-fold and 15-fold validation methods.
5	Bhakti G. Deshmukh et al. (2016)	• Used Granger causality model to observe the association between stock values and twitter sentiments. • Bivariate Correlation test to observe the relation between commodity and actual stock values. • T-paired test to confirm the relation among analyst's recommendations and actual stock values. • Neural Network used for prediction model.	• NSE Companies •Analyst's recommendations, Commodity Prices, tweets (10, 21,000),	• Achieved 70% of prediction accuracy with Twitter sentiments. • Achieved 85% accuracy with commodity prices and tweet generated sentiments • Additional 9% accuracy with analyst recommendations, sentiments and commodity prices.
6	Thien Hai Nguyena et al. (2015)	• The prediction did with SVM model. • Topic-sentiment used to enhance the performance. • Threshold value α is considered for evaluation purpose.	• 18 stocks(Apple, Dell, eBay and so on) • 1-year mood information datasets, specialized features, Historical data.	• Prediction % Accuracy achieved based on 6 features is 52.34, 54.25, 51.87, 52.27, 51.54, 54.41. • At 50% threshold value, achieved more accuracy of 3.03 % with human sentiment and 9.83 % more with aspect-based sentiment model.
7	Kin-Yip Ho and Wanbin (Walter) Wang (2016)	• Used Artificial Neural Network approach for prediction. • Calculated Daily Sentiment Score using the Relevance Score and ESS.	• Google Stock. • 3 years News dataset, Historical data.	• Achieved 54.47 % prediction rate of PNN model. • Reported sensitivity value - 52.83 % and the specificity value - 55.71 %.
8	Tong et al. (2017)	• Used Logistic Regression and SVM models as baseline models. • Conducted Correlation and Causality tests to find relation among different variables. • Used NBSVM, LSTM and Ensemble method for classification.	• China stock market. • Data from microblogger the site, and web forums (8 million posts), Historical data.	• Found powerful correlation and causality among chat room post sentiment score and stock price. • Achieved prediction accuracy of 57.3 • Profit gain of 19.54% at the end of the seven months.

9	Vaishali Ingle et al. (2016)	• The Hidden Markov Model along with extracted term frequency – inverse term frequency features to predict one day ahead stock market value. • Viterbi path calculated.	• 10 companies from NSE • News data of the last 7 days, Historical data.	• Achieved error rate ranging between 0.2 to 3.9 % for weekly collected stock data with an actual closing price and predicted closing price.
10	Rakhi Batra, Sher Muhammad Daudpota (2018)	• Used SVM Model for Stock prediction. • Extracted tweets with sentiments using pipeline API.	• Apple stock. • 7-year tweets, Historical data.	• Results revealed the existence of positive relation among people opinion and market ata. • Implemented work yielded an accuracy of 76.65 % for stock prediction.

B. Big Data Framework and Machine Learning Techniques

The following section addressing the studies analyzing a huge quantity of data generated by social media events, news data and other data sources with the use of big data analytics. The authors in [11] implemented a Naive Bayes algorithm for forecasting stock prices of data gathered from social media events in distributed Map-Reduce programming model. Two input datasets for varied time intervals (5 minutes, 90 minutes, 15 minutes) were used for the prediction task. For estimating the new close price of a stock, classification results obtained from tweets sentiment analysis combined with past prices using Linear Regression. Further, they made a comparison between the calculated new price and the actual stock price using Mean Square Error (MSE).

The authors in [12] used optimized technical parameters with past data to forecast the stock market. The authors utilized Big data Apache Spark platform to implement Genetic Algorithms and further optimized parameters were passed as input to the Neural Network to make buy/sell/hold predictions for the Dow Stock Exchange. Their achieved results have shown that optimized technical indicator parameters enhanced the stock trading performance and provided a strategy for buy/hold decision.

The authors in [13] used financial news and Twitter data to make future predictions about stock values and correlation was used to find the relationship between stock values and sentiments values derived from news, tweets. They have developed the model to make predictions in real time, using Big Data Analytics by applying machine learning algorithms [33, 34]. Their findings show that political news, economic factors, and social media content fully affect the future performance of the system and uncertainty of the market. They also revealed that prediction quality may be improved using numeric data with social media obtained data.

The authors in [14] proposed a Cloud-Based Stock forecasting model by implementing a Neural Network on the Hadoop platform. Along with historical data, they have used a simple moving average, on balance volume and exponential moving average. They achieved finite results in a limited duration of time and enhanced the performance of BPNN by implementing Map Reduce procedure.

The authors in [15] proposed two prediction models supported news, tweets, and historical price. One model for online mode and another for offline mode. The online or real-time model used to forecast during market operating hours and offline model works with historical data including today's market data also. They designed the hybrid model with Big Data Spark framework and HDFS to handle a large

The authors in [16] proposed a way for finding the polarity scores using the classifier Naive Bayes followed by system implementation with neural network stock price prediction by considering historical stock data and sentiment scores as input. They used Hive ecosystem for storage and pre-processing of data. Their observations have shown that the model worked best by providing recent data as input and accuracy reached 98% and proved that the system will generate more accuracy with recent data. Furthermore, the proposed model is employed to determine a statistical link among historic numerical data and other sentimental factors which may cause stock market fluctuations.

The authors in [17] utilized trading price, trading volume, high, low, open price and quantity of selling stocks as input to a neural network to predict stock exchange. They used matching patterns in past stock data to attain everyday stock values and solid rules to choose the major factors that considerably have an effect on the value. They proposed a completely new approach that discovers the best historical dataset with matching patterns supported some machine learning methods. They utilized the Dynamic Time Warping (DTW) algorithm to discover patterns with the utmost likely state matching the existing pattern followed by a stepwise regression analysis to pick out the foremost impacting determinants for the stock price. The prediction accuracy tested with Jaro-Winkler distance values. Table II shows detailed summary of studies based on Big Data Analytics.

Table II Summary of studies based on Big data Analytics

S.No	Author Name and Year	Methodology	Forecasted Index and Input datasets/variables	Results
1	Michał Skuza et al. (2015)	• Used Map Reduce programming model. • SentiWordNet dictionary used for classifying tweets using Naive Bayes classifier. • Linear Regression for stock price prediction with 2 datasets for distinct time intervals.	• Apple Stock. • Twitter data over 3 months (15 million tweets) and Historical prices.	• The model shows the relationship among the information about social services and the stock market. • Forecasting of stock values depend mainly on the selection of training dataset, the number of posts appeared per time interval. • Predicted and actual stock prices are compared with mean square error method.
2	Omer, Murat et al. (2017)	• Used Apache Spark platform. • Genetic Algorithms to improve RSI parameters. • Used Deep Neural Network to make predictions (buy/sell/hold).	•DJIA. • Historical prices, technical parameters named William % R, simple moving average, relative strength index.	• 71.63% of success was achieved for proposed model. • Evaluated success rate 70.88% for Genetic algorithm.
3	Girija Attigeri et al. (2015)	• Technical analysis performed with historical data. • The fundamental analysis performed with social media data. • Logistic regression used for prediction purpose. • Used Hive Script for sentiment analysis on Hadoop platform.	• NSE Stocks. • 355 news articles & 430 tweets for one company. • 510 news articles & 599 tweets for other Company.	• Reported that political news and social media events may cause the the unpredictability of the markets. • The assistant of big data technology allows predictions to make it in a real-time.
4	Kushagra et al. (2013)	• Back Propagation Neural Network Automated approach.	• NSE Stocks. • Historical data, simple moving average, on balance volume, exponential moving average.	• Obtained more prediction accuracy by enabling parallel processing of input data using Hadoop.
5	Mostafa et al. (2018)	• Build the hybrid model that works for real-time and offline mode using Apache Spark and Hadoop HDFS. • Classifier naive bayes for sentiment analysis. • Multiple classifiers used for testing the proposed model.	•Apple, IBM, Google Stocks. •News, tweets, days return, multiple day's returns, returns moving average.	•Achieved prediction rates for real-time Mode ranging between 70.32 % to 75.8%. • Achieved prediction rates for offline Mode ranging between 86.52 % to 88.63%.
6	Malav Shastri et al. (2019)	• A neural network with two data inputs (3-year data and 1-year data). • Hive framework used for cleaning of data in the Hadoop ecosystem. • Naive Bayes classifier for sentiment analysis.	• Apple Stock. • News, historical data.	• Obtained the accuracy of 91 % for 3 years data as input and 98% for recent data for only one year. • Concluded that stock prices predictions are more effective for a shorter period of time.
7	Seungwoo Jeona et al. (2018)	• Data extraction and data aggregation. •Pattern matching based on DTW • Prediction using ANN with Hadoop and R. •Validation using Jaro-Winkler distance values.	• Hyundai, KIA, and Samsung • Trading price, trading volume, high, low, open. • Price and amount of selling stocks. • Price and amount of buying stock.	•Implemented automatic model to predict the stock price. • Forecasted next day's stock price.

All the papers summarized within the above tables are favoring the prediction return rate or error rate as the performance measure. Some papers also discovered the factors that might directly influence the stock market rates.

III. HOW WE ARE FORECASTING?

A. Predictive Power of Google Trends

Google trend data is effectively to see what is trending or see how many times or how often people search for certain terms. Since trends present the volume and search for distinct keywords, researchers used this data to make stock market predictions. Various studies[28,29] and researches [24, 25, 26,27]show that Google trends help to make better decisions for the stock market and the authors proved a positive correlation between Google trends and market movements.

B. Proposed Methodology

As discussed in the above-detailed study, input parameters provided to any machine learning algorithm play a very crucial role to forecast the stock values. Based on the findings claimed in [15], the prediction accuracy for the stock market achieved by top authors I is 70.21% using Logistic regression. So there seems a scope for improvement. In this work, a new framework for trend prediction is proposed using logistic regression in Big Data Environment. Google trends are integrated with sentiments and historical data and this integrated data is given as input to the prediction model.

Table III Correlation and P- value of stock price with Google Trend data

Google Trend Data	Correlation	P Value
TCS	0.141	0.04
HDFC Bank	0.418	1.79E-06

Table III showing a positive correlation of Google trend data with HDFC Bank and TCS stock values and p-value for TCS and HDFC bank with GT data is within machine precision indicating that the relationship is statistically very significant, so it should work as a good predictor for stock market returns. Trends are collected using the name of the company itself which shows the interest of users in these partic Decer

Fig 1. shows the proposed framework for price short term stock prediction. All components of the model are described as follows

➢ Data Collection: - For experimentation purpose, we have collected qualitative as well as quantitative data. We used twitter messages, news headlines, Google trends, and historical data for the last six months of 2018.

- Qualitative data: - Tweets have been collected from the Twitter website using Rapidminer process [31] and news data using XML API in MS-Excel. We have collected around 325000 tweets and 15000 news headlines for two different Sectors named Banking and Information Technology (1 July 2018 to 31 December 2018).
- Quantitative data: - Historical data is collected from Yahoo Finance [30] and Google trend data is downloaded for HDFC Bank and TCS Company from Google website [32] by selecting Finance category as and India country.

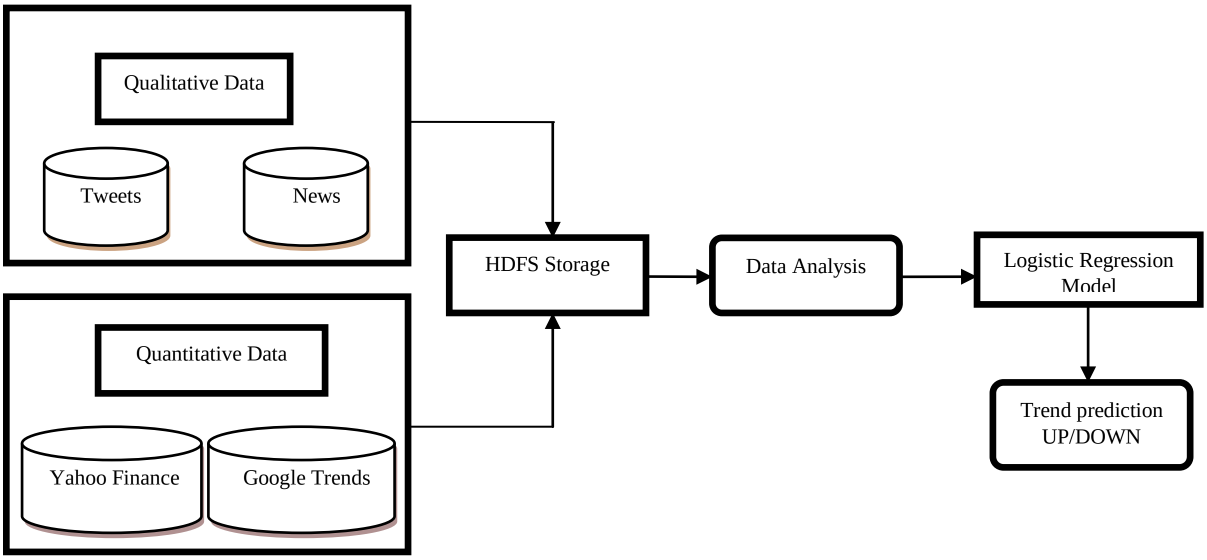

Fig 1. Proposed Framework for short term stock prediction

➢ **Data Storage**: - All the collected raw data is stored in a dispersed manner in HDFS (Hadoop Distributed File System).

➢**Data Analysis**: - Analysis of data has been done using the following steps.

- Data Preprocessing: - Tweets and news that are collected are not usable at this stage for sentiment analysis. The tweets are available as text data and pre-processing step of tweets includes removal ofhashtags, removal of urls and removal of re-tweets and removal of other symbols that may cause inefficiencies in results and tend to decrease the accuracy of the overall process. Pre-processing steps of news include data cleaning (fetching HDFC Bank and TCS news headlines), tokenization and stop word removal.
- Sentiment Analysis: - To extract the sentiment class of tweet text and news headlines, Naive Bayes classifier is empowered with an English lexical dictionary SentiWordNet.

➢ Logistic Regression Model (LR): - We utilized the LR model to find the association among binary class labels (1 or 0) and various features (positive tweet sentiments, positive news sentiments, Google trends, and historical data). The LR model [27] provides the trend probability for a day as 1 (Up) or 0(Down) signals. Inbuilt Logistic Regression function provide by Scala library is used for making the prediction.

➢Trend Prediction: - The implemented model has predicted the future stock trend as 1 or 0.

Table IV shows the Big Data environment that has developed for the experimentation and Fig 2. representing algorithm for Logistic regression implemented in Apache Spark Framework with Scala language.

Table IV Developed Big Data Environment

COMPONENTS	ROLES
Operating System	Use of Hadoop for distributed storage Supporting Java environment for processing Scala logics
HDFS Layer	Storage of data in a distributed manner
SCALA and SPARK Computing System	Sentiment analysis, Text Processing, Algorithm implementation
HIVE	To store analyzed results in HIVE tables and for implementing logics.
Web Server	Analysis of components of Apache Spark

Step 1. Load the data and continue to exist it in memory.

```
val df = spark.read.format("com.databricks.spark.csv").option("header", "true").
option("inferSchema", "true").load("/rs/hdfs_train.csv").
persist(StorageLevel.MEMORY_AND_DISK)
```

Step 2. Create an array of feature columns that would used to predict the trend.

```
val featureCols = Array("positive_tweets_count","close","open","low","high",
"positive_news_count", "google_trends").
```

Step 3. Create an Assembler for the columns created in step 2 and Encode string column

```
val assembler = new VectorAssembler().setInputCols(featureCols).setOutputCol("features")
val df2 = assembler.transform(df)

val labelIndexer = new StringIndexer().setInputCol("trend").setOutputCol("label")
val df3 = labelIndexer.fit(df2).transform(df2)
```

Step 4. Build the Logistic Regression model for the training dataset.

```
val model = new LogisticRegression().fit(df3)
val predictions = model.transform(df3)
```

Step 5. Predict the result of the tested dataset with trained model

```
data set.val xyz=predictions.select ("features", "label",
"prediction").withColumn("flag",when($"label"===$"prediction",lit(1)).otherwise(lit("0")))
val result=xyz.filter($"flag"===1).count.toDouble
```

Step 6. Print the percentage accuracy for tested data

```
val total=xyz.count.toDouble
val accuracy = result/total * 100
```

Fig 2. Algorithm to implement Logistic Regression in Scala

IV. RESULTS AND DISCUSSION

The prediction model has been implemented in Apache Hadoop-Spark environment. Table V representing the prediction accuracy achieved for TCS and HDFC Bank with and without considering Google trends data with sentiment values. Table VI representing pre processing time for tweets and news data of TCS and HDFC bank. Fig 3. shows the prediction accuracy achieved by logistic regression model implemented with available data. The Graph depicts the percentage accuracy for HDFC Bank and TCS by considering Google trends with sentiments and historical data and without considering Google trends with sentiments and historical data. Fig 4. representing the pre-processing time (in seconds) of tweets and news headlines for HDFC Bank and TCS.

Table V Prediction accuracy for TCS and HDFC bank

Company Name	Prediction Accuracy(%) with Google trends	Prediction Accuracy(%) with Google trends
TCS	72.4	68.23
HDFC Bank	74.38	71.07

Table VI Pre-processing time for Tweets and News

Company Name	Tweets(seconds)	News(seconds)
TCS	3300	160
HDFC Bank	3000	130

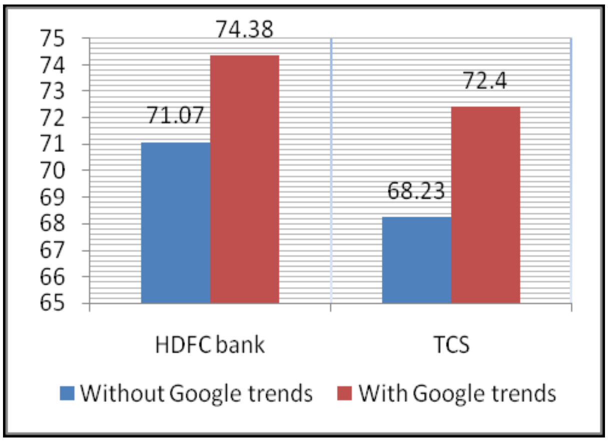

Fig 3. Prediction accuracy of TCS and HDFC Bank

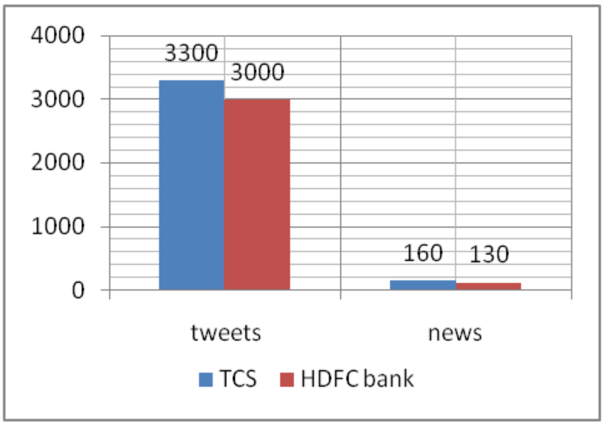

Fig 4. Pre -Processing time (seconds) for tweets and news

V. CONCLUSION AND FUTURE WORK

The stock price trend prediction is complex due to uncertain and volatile behaviour. The selected papers for reviews study making use of collected data from different sources and providing this data as input to the machine learning model creating financial time series forecasting. Additionally, a various number of feasible approaches and the number of ways in which the data can be fed to a machine learning algorithm have been explored for the prediction task. To sum up, review papers suggest that the addition of more input variables for machine learning algorithms yield better results. The proposed model has also improved the prediction accuracy of the Logistic regression model about 3-4 % with the utilization of Google trend data in combination with sentiments and historical data. So the implemented model met with the expectation of above reviewed studies results. This work is conducted as a part of research work and future work involves the implementation of other machine learning algorithms for the stock prediction task in Big Data Environment with a huge amount of data.

REFERENCES

1. D. K. Kirange, Ratnadeep R. Deshmukh," Sentiment Analysis of News Headlines for Stock Price Prediction, " International Journal of advanced computer technology(COMPUSOFT), vol. 5, issue-3,2016.
2. V. Kalyanaraman, S. Kazi, et al." Sentiment Analysis on News Articles for Stock ", 8th Asia Modelling Symposium Kalyanaraman, 2014.
3. D. Yan et al." Predicting Stock Using Microblog Moods," Journals & Magazines, China Communications, vol.13, issue- 8, 2013.
4. K. Joshi, Bharathi H. N, and J. Rao, "Stock Trend Prediction using News Sentiment Analysis, "International Journal of Computer Science & Information Technology (IJCSIT), vol. 8, no. 3,2016.
5. B. G. Deshmukh, P. S. Jain et al. "Spin-offs in Indian Stock Market owing to Twitter Sentiments, Commodity Prices and Analyst Recommendations", International Conference on Advances in Information Communication Technology & Computing,2016, pp. 67-76.
6. T. H Nguyen, K. Shirai, and J. Velcin, "Sentiment Analysis on Social Media for Stock Movement Prediction, " Expert Systems with Applications, vol. 42, issue 24, pp. 9603-9611, 2015.
7. K. Y. Ho and W. Wang, Predicting Stock Price Movements with News Sentiment: An Artificial Neural Network Approach Artificial Neural Network Modelling, pp 395-403, 2016.
8. T. Sun, J. Wang, et al." Predicting Stock Price Returns using Microblog Sentiment for Chinese Stock Market", International Conference on Big Data Computing and Communications, 2017.

9. V. Ingle and S.Deshmukh, "Hidden Markov Model Implementation for Prediction of Stock Prices with TF-IDF features," International Conference on Advances in Information Communication Technology & Computing, Chengdu, 2016, pp. 87-96.
10. R. Batra and S.M. Daudpota "Integrating Stocktwits with Sentiment Analysis for better Prediction of Stock Price Movement," International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), Sukkur, 2018, pp. 1-5.
11. M. Skuza, A. Romanowski "Sentiment Analysis of Twitter Data within Big Data Distributed Environment for Stock Prediction," Federated Conference on Computer Science and Information Systems (FedCSIS), Lodz, 2015, pp. 1349-1354.
12. O. B. Sezer, M. Ozbayoglu and E. Dogdub, "A Deep Neural-Network-Based Stock Trading System Based on Evolutionary Optimized Technical Analysis Parameters," Procedia Computer Science, vol. 114, pp. 473-480, 2017.
13. G. V. Attigeri, M. P. M M et al." Stock Market Prediction: A Big Data Approach," IEEE Region 10 Conference, Macao, 2015, pp. 1-5.
14. K. Sahu, R. Pawar et al." Stock Exchange Iforecasting using Hadoop Map-Reduce Technique, "International Journal of Advancements in Research & Technology, vol. 2, issue 4, 2013.
15. M. M. Seif, E. M. Ramzy Hamed, and A.E. F. Abdel Ghfar Hegazy, "Stock Market Real Time Recommender Model Using Apache Spark Framework", International Conference on Advanced Machine Learning Technologies and Applications (AMLTA), 2018, pp. 671-683.
16. M. Shastri, S. Roy, and M. Mittal, "Stock Price Prediction using Artificial Neural Model: An Application of Big Data," EAI Endorsed Transactions on Scalable Information Systems, 2019, pp.671-683.
17. S. Jeona and V. Chang, "Pattern Graph Tracking-based Stock Price Prediction using Big Data, "Future Generation Computer Systems, vol 80, pp.171-187,2018.
18. G. A. A. Jabbar Alkubaisi, S. S. Kamaruddin and H. Husni, "Conceptual Framework for Stock Market Classification Model Using Sentiment Analysis on Twitter Based on Hybrid Naïve Bayes Classifiers ", International Journal of Engineering and Technology, pp. 57-61, 2018.
19. V. Rajput, S.Bobde," Stock Market Forecasting Techniques: Literature Survey," International Journal of Computer Science and Mobile Computing, vol. 5, pp.500 – 506, 2016.
20. B. Weng, M. A. Ahmed and F. M. Megahed, "Stock Market One-Day Ahead Movement Prediction Using Disparate Data Sources, "Expert Systems With Applications, vol.79, pp. 153-163,2017.
21. S.F. Shazmeen, M. M. Ali Baig and M.R. Pawar," Performance Evaluation of Different Data Mining Classification Algorithm and Predictive Analysis", IOSR Journal of Computer Engineering, vol 10, Issue 6, pp. 01-06, 2013.
22. Pujana Paliyawan, "Stock Market Direction Prediction using Data Mining Classification, "ARPN Journal of Engineering and Applied Sciences, vol 10, issue. 3, 2015.
23. S.Prasanna, D. Ezhilmaran, "An Analysis on Stock Market Prediction using Data Mining Techniques", International Journal of Computer Science & Engineering Technology, "Sudan, 2013.
24. D. Challet, A. B. Hadj Ayed, "Predicting Financial Markets with Google Trends and Not so Random Keywords," SSRN Electronic Journal · July 2013, DOI: 10.2139/ssrn.2310621
25. P. F. Pai, L. Chuang Hong, and K. Ping Lin, "Using Internet Search Trends and Historical Trading Data for Predicting Stock Markets by the Least Squares Support Vector Regression Model," Computational Intelligence and Neuroscience, vol. 2018, https://doi.org/10.1155/2018/6305246
26. F. Ahmed, S. Hina et al. "Financial Market Prediction using Google Trends," (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 8, No.7, 2017.
27. H .Gunduz, Z. Cataltepe and Y.Yaslan," Stock Daily Return Prediction using Expanded Features and Feature Selection, "Turkish Journal of Electrical Engineering & Computer Sciences, vol. 25, pp. 4829 – 4840,2017.
28. https://www.mattmoocar.me/AppleStockPred
29. https://thetradable.com/how-to-predict-the-market-with-google-trends
30. https://Yahoofinance.com
31. https://my.rapidminer.com/nexus/account/index.html
32. https://trends.google.com/trends

AUTHORS PROFILE

Sneh Kalra is a research scholar at MVN University. She has achieved her M.Tech Degree in Information Technology and B.Tech Degree in Computer Science. Her research interests include Data Mining, Big Data Analytics and Machine Learning. She has published 5 papers in international journals and international conferences.

Dr. Sachin Gupta, Dean, School of Engineering and Technology at MVN University has over 16 years academic and industrial experience. His research interests include Blockchain, Wireless sensor networks, Social Media mining and Security. He has published 5 papers in National and international journals of repute, and has contributed 10 books/chapters in Computer Science domain.

Dr. Jay Shankar Prasad is a professor at KEC Ghaziabad. His research focuses on applications of AI and machine learning techniques, wireless networks path optimization, Computer Vision issues, to apply the soft computing techniques to control the Humanoid Robots, Gesture Recognition, ISL Recognition, Pattern mining etc. He has published 22 papers in International journals and International conferences. He has 17 years of teaching and 3 years of software industry experience. He also guided many doctoral, postgraduate and undergraduate level projects.

Retrieval Number: I8474078919/19©BEIESP\
DOI:10.35940/ijitee.I8474.078919

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication