Received: April 15, 2020. Revised: May 14, 2020.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

An Efficient Stock Market Trend Prediction Using the Real-Time Stock Technical Data and Stock Social Media Data

Lakshmana Phaneendra Maguluri[1]* **Ragupathy Rengaswamy[1]**

[1]*Department of Computer Science and Engineering*
[1]*Annamalai University, Annamalai Nagar, Chidambaram, Tamil Nadu 608002, India*
* Corresponding author's Email: phanendra51@gmail.com

Abstract: Stock market trend prediction is one of the major issues which have been problematic to both financial analysts and scientific research on real-time streaming `data. Also, forecasting stock market returns based on the technical data is difficult due to noise in the data and mixed data types. As the size of stock technical tools and techniques are increasing along with stock news data, it becomes more difficult to analyze and predict the market trend based on the historical market data. Therefore, it is very challenging task to research institutes and financial brokerages for data analysis and market trend prediction. Since, financial data is unstructured by nature and it is difficult to find an essential feature for intraday bullish and bearish trend prediction. In this work, a new stock technical indicator based non-linear SVM model is designed and implemented on the real-time stock market data for trend prediction. In this model, a novel stock technical data transformation technique, stock technical and text feature extraction method and non-linear SVM classification algorithm are proposed to predict the stock trend on daily and weekly basis. Experimental results have shown that the proposed stock market trend prediction approach has 7% of computational runtime (ms) and an average stock prediction accuracy of 10% as compared to the existing stock market trend prediction models.

Keywords: Stock market prediction, Support vector machine, Sentiment prediction, Trend analysis.

1. Introduction

Stock market is considered as the most vital and active part of financial institutions and investors. The financial articles and trend data have been considered as the primary factor for market trend prediction. Most of the organizations are depend on the high computational systems in order to predict the market trend based on the sentiment score and stock technical data. These predictions are used to filter positive and negative sentiment stocks are been used to take appropriate decisions by the investors. Hence, the modelling and analysis process of news articles are very much essential in order to make accurate predictions. The organizations can become market prominent, if they can attract the attention of more and more investors [1]. Let us consider an example, if an investor has 100 stocks; though the investor couldn't able to filter the feasible top positive and negative trend stocks based on the company announcements and technical data as shown in Fig. 1.

Stock exchanges are inconsistent, it may vary over time; the correlation between social media and

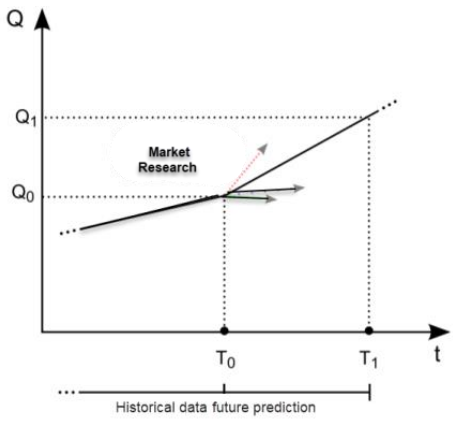

Figure. 1 Historical stock market trend prediction

stock market data certainly varies over time. A large number of methods have been proposed in the literature to find and predict the correlation of the stock features with financial time series data. These methods are applicable to structured and numerical databases. In the traditional systems, different market analytical tools and statistical analysis tools are used to predict the sentiment of the stock market trend. In the stock market, the share prices can depend on many factors, which are ranging from corporate news to political news [2].

Stock exchanges can be seen as systems that vary over time; the correlation between social media and stock market data varies certainly over time [3]. Taking time varying behaviour into account may result in more accuracy as correlations that vary in time can now be taken into account. This is particularly useful in the prediction of stock markets using social media information because no time-invariant correlation can be anticipated. For example, a credit crunch search term may not be equally relevant for months [4]. Assuming that a constant correlation for a long time would probably lead to incorrect consideration of the data when it is no longer useful or exclude data for time periods for which useful information may be provided. Really variable SVR extensions are not known. To implement time-diverging behaviour into the SVR, the horizontal approach is applied. The SVR is trained on a limited number of previous data points each time, only to focus on the latest history. This method is easy to implement and was used previously with the LASSO [5]. The basic formulation of the LASSO shown in Eq. (1). In the Eq. (1), A is the input data matrix, x,y are the variables, ε is small constant, E is the expectation of the variable. J is the predicted value.

$$J = \|Ax - y\|_2^2 = x^T(A^TA)x - y^TAx - x^TA^Ty + y^Ty$$

$$\frac{\delta J}{\delta x} = 2(A^TA)x - 2A^Ty = 0 \ \textit{(solution)}$$

$$x = (A^TA)^{-1}A^Ty$$

$$E[x] = E[(A^TA)^{-1}A^T(Ax + \epsilon)] = E[x] + E[\epsilon] = x$$

$$\Sigma_x = E[x^2] - E[x]^2 = E[((A^TA)^{-1}A^T(Ax + \varepsilon))((A^TA)^{-1}A^T(Ax + \varepsilon))^T] - xx$$

$$= E[(x + (A^TA)^{-1}A^T\varepsilon))(x + (A^TA)^{-1}A^T\varepsilon))^T - xx^T \quad (1)$$

Stock markets have been studied repeatedly to develop useful models and to forecast their movements. Most of the researchers and financing investors use statistical tools and techniques in order to predict stock market conditions. The behaviour of the stock market is generally not known to financial analysts who invest in stock markets [6]. They can immediately act on it, if they can forecast the future action of stock prices. Pricing trends based solely on data analysis are very popular. However, only the event is included in numerical time series data and not the reason. The use of textual information, in combination with numerical time-series data increases the quality of input and improves stock prediction. The application of these methods establishes the link between news and stock prices and allows us to learn a prediction system using a text classifier. The technical analysis, which uses technological analytical indicators, is built on the numerical time series data and attempts to predict stock market trend. It is based on the hypothesis that all the reactions to news on the market in real time are included to predict the market future price action. The main objective of this work is to identify current trends and predict future stock trends in charts [7]. In chart analysis, the timing of the market is considered critical and opportunities are identified by estimating historical changes in prices and volumes and comparing them with current prices. Technicians charts and models are used to determine trends in price and volume. Different types of Web-based financial information sources provide electronic versions of daily issues. All these sources contain political and economic global and regional news, quotes from leading bankers and politicians and financial analysts ' recommendations [8].

Technical analysis with historical trading data can be used for data analysis. This includes inventory, future, commodity, fixed income, currency and other securities. Technical analysis is much more prevalent in commodities and foreign exchange markets, where traders concentrate on short-term price movements [9].

Technical indicators, known as' technical' ones, focus rather than on the fundamentals of a business, such as income, revenue and profit margins, on traditional trading data such as price, volume and open interest. Technical indicators are commonly used by active traders because they are intended to analyze short-term price moves, but long-term investors can also use technical indicators for identification of points of entry and exit.

Relative Strength Index: Relative Strength index

Figure. 2: Basic Sentiment based Classification model for Trend prediction

is a measure of the strength inherent in a field calculated using the amount of price changes upwards and downwards over a given time frame. It has a range of 0 to 100 with typically 30 to 70 values. Higher Relative Strength Index values indicate overbooked conditions while lower values indicate oversold conditions. The formula is as follows for the calculation of the relative strength index shown in Eq. (2)

$$RSI = 100 - \frac{100}{1 + RS} \quad (2)$$

Where as $RS = Average\ Gain\ /\ Average\ Loss$.

RSI is the relative strength index which is used to find the trend of the stock based on the candlestick values.

Money Flow Index: The Money Flow Index is a measure of the strength of the currency instrument that flows into or out of an open-market stock. It is mainly derived by comparing the volume of upward and downward changes in prices over a given period. The Cash Flow Index is based on the amount of the Cash Ratio, which is the ratio of positive cash flow to negative cash flow over the period concerned [10].

Moving average: The move average of a field is returned for a given period of time. The moving average is calculated by combining past values with current values over the given period.

MACD: The MACD is the difference between the short and long moving field averages. The MACD is usually a particular instance of a value oscillator and is used mostly to detect price trends at the closing price of a security system. If the MACD is on a growing trend, prices are higher. If the MACD is on a downward trend, prices are lower [11].

Many feature selection metrics have been examined for text categorizations, including information gain (IG), chi-square (CHI), coefficient

Feature subset selection can improve the classification accuracy by creating optimal subset features from the high dimensional feature space. A feature selection approach has been used for the selection of limited features from the original stock feature selection [13]. Adaboost(Adaptive Boosting) of correlation (CC) and odds ratios (OR). An overview of the comparative study of various functional selection metrics demonstrated that the conventional selection scores are still the best for categorization of text.

The main contribution of the work is to find the sentimental based stock classification in order to improve the buying behaviour of investors on the Futures market as shown in Fig. 2. As shown in figure, stock news and real-time data are taken as input for sentiment analysis and technical analysis. These features are given to classification model for stock trend prediction. Sentiment analysis reveals the effect of unstructured market data on investor emotions for decision making. The market feelings or the purchasing behaviour of the merchants are based on the stock technical and the market trend.

This work is an extension to hybrid Bayesian network-based stock market prediction model [12]. In this work, a hybrid Bayesian classifier is used to predict the stock trend for real-time dataset. In this paper, a novel real-time stock market trend prediction by using data pre-processing and non-linear classifier. This model predicts the stock trend in time basis.

The main advantage of the proposed model is to find and predict the trend of the stock in time basis. This model is better applicable to real-time stock trend prediction in intraday and customized time manner. In this paper, a novel technical indicator is proposed to predict the trend of the stock using the training data.

The rest of the paper was organized as follows: Section 2. Deals with Related work; Section 3. Filter based stock technical prediction model; Section 4. Conveys the obtained results; Section 5. Inference; Section 6. Discuss about the Conclusions drawn.

2. Related works

is a meta learning based approach from the ensemble learning group. The main objective of the AdaBoost is to improve the strong classifier using the group of base weak classifiers. AdaoBoost approach is an iterative method, and in each iteration, a weak base classifier is selected to

minimize the error rate of the model. High dimensionality is one of the severe issue for machine learning models In order to optimize the precision of the classification algorithm, most of the classification approaches use feature selection measures such as mutual information, correlation coefficient, rough-set, chi-square test etc.,[14]. To select subset of features from the high dimensional space. They implemented a PSO based spectral filtering model to high dimensional features of the original training data. Two reconstruction methods are used, one is the principle component analysis and the other is Maximum likelihood estimation. Several distribution algorithms were used in randomization models. In most of these approaches Bayesian analysis is used to predict the original data distribution using the randomization operator and the randomization data [15].

A multi-scale filter bank is used in order to present the characteristics of stock trend image data texture and structure. Different efficient and effective classification schemes are implemented to train the system. In the subsequent time, another generalize system is developed which has the responsibility of regional trend classification. The basic prior probability of the stock trend prediction as shown in Eq. (3). Here X is stock random variable, C is stock class label, N total number of stocks, τ is the scaling factor (0.5), $\emptyset$ is the parametric equation.

$$\begin{aligned} Pr(X < C|X < 0) &= Pr(y + \frac{\tau}{2N} < C) \\ &= Pr(N\left(X^{true}, 1/\sqrt{N}\right) < C - \frac{\tau}{2N}) \\ &= Pr(N\left(0,1/\sqrt{N}\right) < C - X^{true} - \frac{\tau}{2N}) \\ &= \sqrt{N}Pr(N(0,1) < \sqrt{N}(C - X^{true}) - \frac{\tau}{\sqrt[2]{N}}) \\ &= \sqrt{N}\emptyset(\sqrt{N}(C - X^{true}) - \frac{\tau}{\sqrt[2]{N}}) \end{aligned} \quad (3)$$

Most of the traditional approaches detect inappropriate and computationally infeasible patterns on high dimensional datasets. Hence, it is difficult to process all of the stock patterns that are not required during the process of classification. Hence, the overall computational overhead also increases significantly. Unwanted noise is resulted during the process of classification. Hence, it is very much required to select essential stock patches during the classification process. All of the traditional stock selection techniques involve a perfect combination of filter and wrapper schemes. Filtering approaches have the responsibility to rank every individual feature according to their goodness. During the process of ranking, the relationship among every individual stock with respective class label is considered. Univariate scoring metric play a significant role in the above ranking process. It's important to note that there are other approaches that can be taken to make sure that your causality testing is done properly when the time-series you're using are non-stationary as shown in Eq. (4)

$$\begin{aligned} RSS_{AR} &= \sum_{n=1}^{N}(y_n - \sum_{i=1}^{D^y} x_{y,i}y_{n-i})^2 \\ SS_{ARX} &= \sum_{n=1}^{N}(y_n - \sum_{i=0}^{D^z-1} x_{z,i}Z_{n-i} + \sum_{i=1}^{D^y} x_{y,i}y_{n-i})^2 \\ u &= \frac{(RSS_{AR} - RSS_{ARX})/D^z}{RSS_{ABX}/(N - D^z - D^y)} \sim F_{D^zN-D^z-D^y} \end{aligned} \quad (4)$$

RSS_{AR} is the deviation of the expected and observed values of the stock data(x, y). SS_{ARX} is the sum of squares of deviations of observed and expected values with z-tabulated values. Where N,D are the number of samples and D is degree of freedom.

The top ranked stock candle stick patterns are selected prior to the execution of classification schemes. On the contrary, wrapper schemes require the stock selection approach in order to integrate with a classifier. The prime objective of this technique is to evaluate the classification performance of every individual stock subset. The optimal subset of trend patterns is detected according to the ranking of each feature. Traditional filtering schemes are incapable and inefficient to measure the relationship in between different stocks [16].

Directional Accuracy (DA) and Mean Absolute Percentage Error (MAPE) methods are used to evaluate each feature or feature subset to optimize the classification accuracy [17]. Filter method evaluates each feature independent from the classification algorithm, ranks the stock features after evaluation and considers the superior one. This evaluation is performed using information, dependency, distance and consistency. The basic DA and MAPE functions are shown in Eq. (5)

$$DA_n = 1 \quad if\left(\hat{y}_{n+1} - y_n\right)(y_{n+1} - y_n) > 0$$
$$else\ DA_n = 0$$

$$DA = 1/N \sum_{n=1}^{N} DA_n$$

$$MAPE = 1/N \sum_{t=1}^{n} \left| \frac{y_n - y\hat{}_n}{y_n} \right| \quad (5)$$

Here N is the number of stocks, y is the random variable for data prediction. DA and MAPE are used to estimate the trend of the stock data based on the training values.

In general, the speed of wrapper model is slower than the filter model because of cross validation and repeated iteration to evaluate the feature subsets. Traditional wrapper model is more efficient because classification technique affects the overall accuracy, although the subset selection is an NP-hard. However, if the number of features involved in complex data increases, finding new trend patterns can become difficult due to the complex relationships among features. Feature ranking methods compute the measure for each feature and rank them accordingly. These ranking methods select the top 'k' features based on highest rank and eliminate those having lower feature ranks [18]. Information gain is one of the attribute selection measures which are based on entropy value. Efficient sub-sets of attributes contain class-related and non-related attributes. The method is used to determine how closely the characteristic vectors are linked [19]. When the coefficient of correlation between the two vectors is above "0," the characteristics are said to have a highly positive correlation. Likewise, the functions are said to be negatively correlated if the correlation coefficient between these two characteristic vectors is less than "0" The features are said not to be correlated if the correspondence coefficient of the two characteristic vectors is equal to "0"[20]. Chi-Square is based on the analysis of statistics. The functional vector measures the independence. The strength of the relationship between two random variables is tested with observed and anticipated values. The descriptors should ideally be invariant to operations like scale, rotation and illumination changes. This invariance enables descriptors to be matched across videos which have differences in these parameters.

Extreme classification model is an extension of traditional neural network model for data classification. It partitions the whole problem into numbers of sub-problems and merges them to find an optimal stock market trend prediction. The parameters of hidden layer contain training data samples are mapping to output layer. In the traditional Feed-Forward Neural Networks (FFNN) approach, the adjustments of parameters are iterative in nature and results some issues. These issues are overcome by the suggested Extreme classifier approach [21]. Most of the traditional learning models for training Single Hidden Layer Feed-Forward Neural Networks(SLFN) are comparatively slower than that of non-parametric approaches on stock detection. This approach operates slowly because parameters are required to be tuned iteratively. Moreover, these models require high computational memory and also increase the overall computation time of the mapping process. An extended and slightly modified version of traditional FFNN approach is developed as Extreme classifier [22]. This method is used to enhance the efficiency and performance of conventional SLFNs. Also, most of the Neural network-based learning schemes perform manual tuning of control parameters (such as learning rate, learning epochs etc.) as well as local minima. But Extreme classifier is applied automatically and there is no need of manual iterative tuning. The classification boundary is not optimal in FFNN and the boundary is constant throughout the stock training phase. Hence, there are chances of misclassification of samples closer to boundary. This approach requires a large number of hidden neurons as compared to other traditional tuning-based approaches.

Each stock data is scanned and transformed into normalized continuous data. The main issues of the stock datasets are high dimensionality and imbalance nature. Traditional machine learning classifiers consider subset of features for classification and trend prediction with high true negative rate and error rates. Attribute selection is used to compute the measure for each feature and rank them accordingly. These ranking methods select the top 'k' features based on highest rank and eliminate those having lower feature ranks. Information gain is one of the attribute selections measures which is based on entropy value. Information gain approach is the mutual information of a target random variable say P and independent random variable Q. The main limitation of this approach is, it chooses features having large distinct values over the features having less distinct values. They developed an evolutionary technique in order to detect stock anomalies using hidden Markov models for imbalance detection. In this work, they proposed an imbalance-based technique which is responsible for monitoring the bandwidth consumption of the sub-stock. The normal behaviour scheme completely depends upon the bandwidth consumption of the sub-stock. The most common variables of hidden Markov models are:- bandwidth consumption and the total amount of time required for all stock activities. In this model, a feature ranking measures such as information gain

and correlation are used to filter the feature space. After successful completion of the feature ranking, feature reduction approach is implemented. The feature reduction technique is implemented through the integration of ranks generated from the process of information gain and correlation. The reduced features are given as input to feed forward neural stock to train and test the stock features in stock dataset. In this method, the pre-processing is carried out manually which is a severe drawback of this model.

Mittal, et.al, proposed knowledge-maximized ensemble approach for various kinds of concept drift [23]. In this work, they presented an advanced data stream classifier which is known as knowledge maximized ensemble. Hence, it becomes hard and complicated to restrict the amount of training data. This technique can be influenced by different kinds of concept drift through integration of various imbalance detection approaches. Decision tree induction is a simple and powerful classification process that produces a tree and a set from a specific dataset [24] representing a model for different classes. During regression analysis, the linear combination splitting criteria are used. In building a classification model, the training data set is used, while the test data record is used in model validation. It is used to classify and predict new records, which differ between training and testing. Controlled learning algorithms are preferable to unchecked learning algorithms (like clusters), because their prior knowledge of class labels of data logs simplifies the selection of features / attributes, and therefore leads to the prediction / classification of accuracies. Some researchers have succeeded in adopting the theory of the Rough sets for the classification of different stock complications. The error rates for rough sets were found to be completely comparable and often significantly lower than the other computational techniques.

Recently, ensemble learning models have become popular and widely accepted for high dimensional and imbalanced datasets. Most of the traditional ensemble classification models are processed with limited feature space and small data size. As the size of the feature space increases, traditional ensemble classifiers select a predefined number of features for classification. Learning classification models with all the high dimensional features may result serious issues such as performance and scalability. Feature selection measures can be categorized into three types: wrappers, filters and embedded models. Several studies have been carried out with the aim of classifying respiratory trends with high-classification accuracy using different types of Artificial Neural Network (ANN) architectures developed in different datasets. Random Forests, Gradient Boosting, or even Logistics Regression can also be used to predict and classify trend with high dimensional feature sets. Although high-classification accuracies have been reported, further dynamic evaluation of trend is needed to gain information about the stocks. It is important to measure trend function because stock frequently do not recognize or emphasize small movements, especially if they last longer time. Learning classification models with all the high dimensional features may result serious issues such as performance and scalability [25]. The main problems in the existing models are:

1. Problem of feature selection on high dimensional datasets.
2. Problem of predicting trend with high true positive rate and less error rate
3. Problem of handling high dimensional and large datasets using the parallel processing model.

Class-imbalanced data are common in the domain of data categorization. It generally categorizes many irrelevant documents, but some articles are categorized under interesting category. BN approaches are mostly implemented as standard classifiers. These approaches give rise to exact results along with the capability for representing relationships in between variables. This approach is unable to resolve the traditional class-imbalanced problem [26]. The above process continues executing till it matches the size of other class and cost-sensitive learning scheme. It includes the modifications of relative cost associated with misclassification of positive and negative class [27]. The outcomes of both methods are analyzed and compared with performance achieved without balancing.

3. Filter based stock technical prediction model

In the paper, we have proposed a novel filtered based classification model on the technical data to find the bullish trend stocks on the real-time market data. This model is tested on the continuous type of technical data for trend prediction. In the proposed framework, a novel stock market trend prediction model is designed and implemented on the real-time market data. In this model, real-time stock market technical data and its social medial comments are used to predict the trend of each stock for the classification problem. Fig. 3 describes the

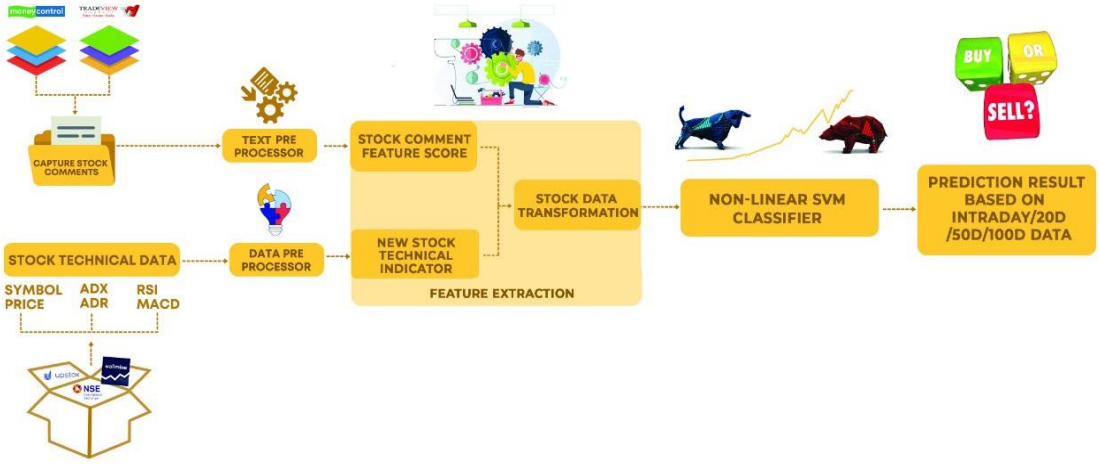

Figure. 3 Proposed stock market trend prediction framework

flowchart of the proposed model for stock market trend prediction. Initially, real-time market data is taken from the stock market sites such as NSE or Upstox etc. Social medial stock comments are extracted from the tradeview or moneycontrol site. Stock related technical factors such as symbol, price, ADX, ADR, RSI, MACD, news sentiment score, etc., are used as the training data. Here, text pre-processing and data pre-processor operations are performance on the stock text comments and technical data. New features are extracted in the comments and technical data for stock trend prediction. Finally, each stock trend is predicted using the proposed non-linear SVM classifier in different levels of time frame.

3.1 Text pre-processor

In a first step all stock market comments of the stocks are captured from the money control website as the training data. All the known stop words are subsequently removed from the corpus. Here we adopt a standard POS tagger for extracting nouns from stock sentiment data. If a model uses only P, then in the left side of Fig. 4 a rational prediction "trends down without reversing" would be used; in the right side of Fig. 4, "trend up without reversing." "Returning and upward trend" in the left side of Fig. 4 and "reversing and downward trend" in the right side of Fig. 4 would increase the model's error rate.
If a model uses only N, derivative f could not explain why the price still "drops down," as the left side of Fig. 4 shows, when good news is released, and why the price is "trending up" when bad news is released.

Figure. 4: Bullish and Bearish trend prediction

In this work, a new comprehensive and finance-specific word-list is used to find the stock sentiment score. Using standard TF / IDF weighting scheme, the weight of each term is used to find the relevance of the bullish or bearish trend in the stock. The weight of the stock comment term t is computed as shown in Eq. (6)

$$w(t,d) = (0.5 + 0.5 \times \frac{f(t,d)}{max_t\, f(t',d)}) \times I(t)$$
$$I(t) = log\frac{C}{c(t)} \quad (6)$$

Where f(t,d) is the frequency of the term in the comment, $max_t\, f(t',d)$ is the maximum frequency of all terms in all the stock comments. C is the count of the both positive and negative classes and c(t) is the maximized positive or negative terms in comments list.

The public sentiment on each stock is specified as bullish and bearish. The bullish of the stock is computed by using following Eq. (7)

$$Bull_s^+ = \sum_{i=0}^{\tau}\sum_{j=0}^{K}\frac{P_{i,j} \times w(t,d)_j}{l_i} \quad (7)$$

A where $P_{i,j}$ represents the count of bullish terms in each comment. Tp_i is the total count of the bullish terms in all the comments , Tn_i is the total count of the bearish words in all the comments, w(t,d) represents the weight of comment, and $N_{i,j}$ represents the number of bearish terms in each comment. The bearish comment sentiment score of each comment is computed as Eq. (8)

$$Bear_s^- = \sum_{i=0}^{\tau} \sum_{j=0}^{K} \frac{N_{i,j} \times w(t,d)_j}{l_i} \times Tn_i \quad (8)$$

3.2 Data pre-processor

In the data pre-processor, each feature in the technical data is normalized to find the correlation between the data samples for trend prediction. Let x defines the input vector x which is normalized in the specified range x ∈ R → x∈ [R1,R2] to remove the sparsity issue. Here R1 and R2 represent the predefined normalized range as shown in Eq. (9)

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \times (R2 - R1) + R1 \quad (9)$$

3.3 Feature extraction

3.3.1. Stock comment feature score

To each comment in the stock corpus D, we construct a dictionary of words that contains bullish and bearish words. By using this dictionary, each stock comment d is represented as a bag-of-words vector w. Let st(i) defines the i^{th} comment term in the stock comment d. In order to compare the comments of different stocks, the term frequencies (tf) are used to normalize the words in all the stock comments. This normalized data is rescaled by using inverse document frequency (idf) as shown in Eq. (10)

$$tfidf(st(i), d, SC) = tf(st(i), d) \times idf(st(i), SC)$$
$$idf(st(i), d, SC) = log \frac{|SC|}{1 + |\{d \in SC | st(i) \in d\}|} \quad (10)$$

where |SC| is the cardinality of stock related comments SC.

3.3.2. New stock technical indicator

Mutual information (MI) is used to find the variation in the two or more data distributions of stock features. In the real-time market data, it is used to analyze contextual feature relationship over time. It is the measure of correlation and dependency of the features in high dimensional dataset. Hybrid Mutual information is represented in terms of bullish and bearish cases as shown in Eq. (11).

$$IG_a(l) = p(l, bu)\, log \frac{p(l, be)}{p(l) \times p(be)} \quad (11)$$

where bu represents the bullish and be represents the bearish type. P(l,bu) is the probability of the term in the bullish dictionary. P(be): probability of the bearish terms. $IG_a(l)$: Information gain o f the term l in the bullish or bearish terms list.

Here, computed stock sentiment score is used as the additional attribute in the training data.

$$I(X,Y) = P_{Joint}(x,y)\, log \left(\frac{P_{Joint}(x,y)}{P_X(x) P_Y(y)} \right) \quad (12)$$
$$l * \left(\frac{x}{\beta}\right) = x^2(w).p(\beta) \sum_{t=1}^{T} I(X,Y)$$
$$for\ technical\ data$$

Let α, β, γ represents the three essential Stock trend technical factors taken from the training data

α = Stock Performance (D)
β = Volatility (D)
γ = RSI (D)

$$Stocksentiment = \frac{(\propto \times \beta)}{\gamma}$$

SSC (3) = Stock sentiment score;

$$P(\beta) = \alpha \times SSC(3) * RSI \sum_{i=1}^{p} \left(\frac{1-\alpha}{2} \beta_j^2 + \alpha |\beta_j| \right), \alpha \in [0,1]$$

Where as RSI>70 oversold
Or RSI<30 undersold

3.3.3. Bullish dictionary words

In this section, a list of bullish dictionary words is used to predict the trend of the stock using the technical indicator. A set of bullish words are described below to compute the trend of a stock in real-time market data.

escalated, gain, enjoy, expansion, aggrandize, elevated, increment, rise, prefer, hallow, expand, supersize, idolize, positive, appreciate, plus,

relish, accelerate, augment, raise, more, amplify, soar, adore, appreciative, approbatory, desire, esteem, approving, raised, swell, extend, addition, worship, climb, add, commendatory, venerate, augmentation, fancy, revere, friendly, proliferate, addendum, increased, escalate, proliferation, accumulate, love, stoke, complimentary, heightened, hype, uprise, accrual, boost, up, applauding, enlarge, admire, admiring, good, multiply, accretion.

3.3.4. Bearish dictionary words

In this section, a list of bearish dictionary words is used to predict the trend of the stock using the technical indicator. A set of bearish words are described below to compute the trend of a stock in real-time market data.

descend, recede, depreciative, abhor, drop, diminution, depreciatory, uncomplimentary, adverse, deplore, slide, detest, lower, plunge, lessen, unappreciative, depletion, dislike, dive, reduce, decrease, depressed, decreased, under, diminish, dip, low, derogatory, disapprove, unfavourable, negative, lowering, loathe, disfavour, unflattering, sink, receded, disdain, hate, decrement, unfriendly, subtract, loss, abate, decline, despise, fall, diminishment, lessening, downsize, abominate, minify, execrate, deprecate, in-appreciative, dropped, shrinkage, reduction, wane, abatement, disapproving, dwindle, down.

3.3.5. Stock data transformation

Input: Training dataset D, F (D): Feature space of D.
Output: Kernel Filtering or Transformed data KD.

Procedure:
Read input data D.
For each pair of feature F[i], F[j] in feature space F (D)
Do
Apply Kernel transformation on I as

$$\text{KernelTransform}(\text{F[i]}) = \emptyset = \frac{1}{\sqrt{2\pi}} e^{-(F[i]-\mu(F[i]))/\sigma(F[i])}$$

Where $n = \sum (F[i] - \mu)/max\{F[i]\}$
If (FT (F[i]>0 and n>0)
Then
Normalize F[i] using Min-max normalization [0, 1]
Else
Normalize F[i] and F[j] within [0, 1] using Min-max normalization as KD

$$x' = \frac{x - min\_(x)}{max\_(x) - min\_(x)} * (R2 - R1) + R1$$

End if
Done

Here, min-max normalization is used to scale the data between the ranges of 0 to 1. This approach is used to clean the data in high dimensional datasets.

Symbol	Notation
F[i]	i^{th} feature
$\emptyset$	Kernel transformation
μ	Mean
σ	Standard deviation
$R1$	Minimum normalized value
$R2$	Maximum normalized value
$min\_(x)$	Minimum value in feature
$max\_(x)$	Maximum value in feature
min_{W_k,a_k}	Minimization of objective function
$\|W_k\|$	SVM parameter
$Ker < x, y >$	Kernel function with x and y values
τ_m, s_i	constants

3.4 Proposed non-linear SVM classification model

Input the stock features for data classification.
For each feature set do
for each stock in SD.
do
Apply SVM multi-class optimization models as $min_{W_k,a_k} \frac{1}{2}\|W_k\|_1{}^2 + \tau_m + \sum_{i=1}^{l} a_i(y_i[Ker < x, y >.w + b] - 1 + s^e{}_i) - \sum_{i=1}^{l} \gamma_i s^e{}_i$

$s.t\ Ker < x, y >.w + b \geq 1 - s_i{}^{en} - \tau_m, s_i{}^{en} > 0, \tau_m > 0; m = 1 \ldots classes$

Here kernel function ker(x,y) represents the kernel functions defined from trend feature space.

$$Ker < x, y >= e^{-s_i^{en\,log(\sum\|x-y\|^2)}} \quad if\ x == y$$
$$= e^{-s_i^{en\,log\left(\sum\|x-y\|^{1/2}\right)}} \quad if\ x < y$$
$$= e^{-s_i^{en\,log(\sum\|y\|^2)}} \quad if\ x > y$$

Test data is predicted to the class y based on the largest decision values as

$$argmax\{W_K^T D_i + b_k\}$$

4. Experimental results

Experimental results are simulated using java environment and real-time market data. In this work, a real-time NSE stock market data are taken as input for technical analysis. These real-time data are taken from the Zerodha/Upstock brokerage API. Proposed model is compared to the traditional stock market classification models to verify the performance of the hybrid classification model to the traditional models. Also, proposed model is compared to the traditional techniques by using various statistical performance measures such as accuracy, true positive rate, recall, precision, false positive rate, ROC area etc. These performance metrics are analysed and compared by using third party java libraries. Different types of statistical metrics such as recall, precision, accuracy, F-measure are evaluated on the stock market sentiment data along with the technical data. These statistical measures are evaluated based on the confusion matrix as shown in Table 1.

Accuracy: It is the ratio of correctly labelled stock predictions class labels to the entire stock class labels as shown in Eq. (13)

$$Stock\ Accuracy\ (SA) = \frac{(STP + STN)}{(STP + SFP + SFN + STN)} \quad (13)$$

Precision: It is the ratio of correctly classified positive stock classes to the all actual positive and negative labelled stock classes as shown in Eq. (14).

$$Stock\ Precision\ (SP) = \frac{STP}{(STP + SFP)} \quad (14)$$

Table 1. Stock measuring metrics

Actual \ Predicted	Stock positive	Stock negative
Stock positive	Stock true positive (STP)	Stock false positive(SFP)
Stock negative	Stock false negative (SFN)	Stock true negative (STN)

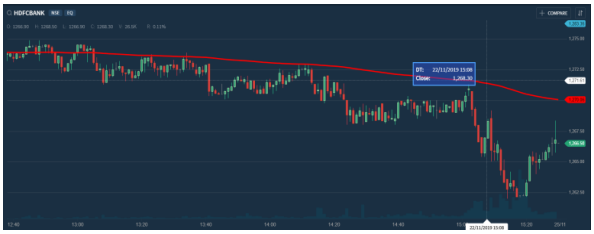

Figure. 6 Upstox 1min candle data for testing

Figure. 7 Upstox 5min candle data for testing

Figure. 8 Upstox one day candle data for testing

Recall: It is the ratio of correctly classified positive stock classes labels to the all predicted positive and negative labelled stock classes as shown in Eq. (15)

$$Stock\ Recall(SR) = \frac{STP}{(STP + SFN)} \quad (15)$$

F-Measure: It is the harmonic average of recall and precision as shown in Eq. (16)

$$Stock\ F - measure(SF) = \frac{2 \times SR \times SP}{(SR + SP)} \quad (16)$$

Fig. 6 illustrates the 1 min candlestick pattern graph in the UPSTOX brokerage website. From the above graph it is clearly noted that the performance of the HDFC bank is UP trend in the afternoon session and neural in the morning session.

Fig. 7 illustrates the 5 min candlestick pattern graph in the UPSTOX brokerage website. From the above graph it is clearly noted that the performance of the HDFC bank is UP trend in the morning session and neural in the afternoon session.

Fig. 8 illustrates the 1 day candlestick pattern graph in the UPSTOX brokerage website. From the above graph it is clearly noted that the performance of the HDFC bank is UP trend up to June month and neutral between July to September months.

Table 2. Stock features extraction using the proposed models

MI [29]	Chi-Square [28]	Rough set [31]	IG [30]	GA [32]	PSO [33]	Proposed FS
58	56	78	63	56	51	41
53	54	78	60	52	56	37
55	54	74	59	62	69	43
56	52	82	57	53	60	46
58	55	63	58	60	72	48
56	59	85	64	54	52	44
65	65	75	58	58	69	42
53	65	77	55	62	75	50
64	63	79	50	62	56	41
65	52	85	59	57	69	36
60	71	69	62	56	50	36
56	61	70	57	61	70	38
59	70	67	65	58	74	42
56	67	69	62	60	58	45
54	59	69	51	55	60	39
51	72	74	60	57	72	45
53	64	61	57	53	60	46
60	59	85	50	58	65	44
56	62	76	62	61	69	38
58	62	71	51	60	54	38

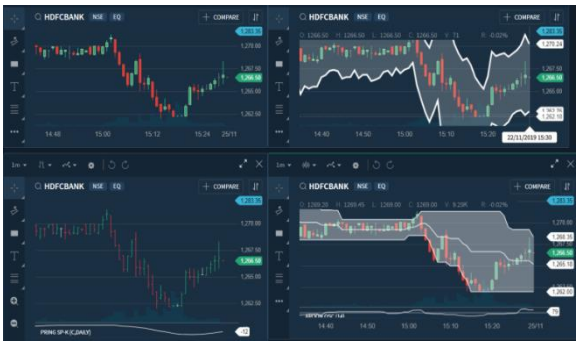

Figure. 9 Multiple technical trend prediction model

Fig. 9 illustrates the various candlestick patterns in the UPSTOX brokerage website. From the above graph it is clearly noted that the various types of technical trends are used on the HDFC bank to check the trend from the morning session to the afternoon session.

Table 2 illustrates the performance of stock trend feature extraction using the proposed approach on large datasets. From the table1, it is clearly shown that the present feature extraction procedure has high filtering rate as compared to the existing approaches.

Fig. 10 describes the sample yes bank user comments in the money control website. This data is extracted using the web drivers in real time.

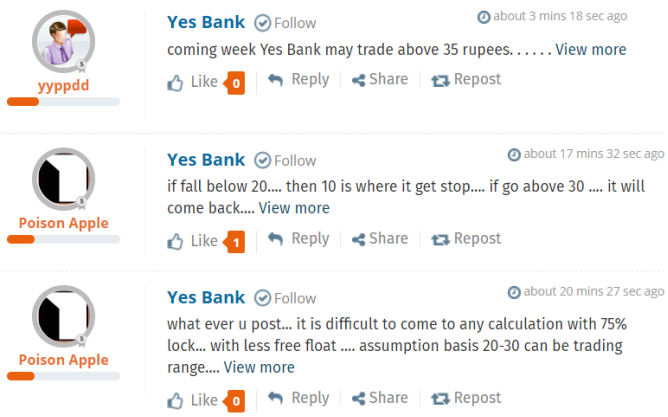

Figure. 10 Sample Yes bank user comments

Table 3 illustrates the performance of stock trend feature extraction using the proposed PSO approach on large datasets. From the Table 2 it is clearly shown that the present feature extraction procedure has high filtering rate as compared to the existing approaches.

Table 4 describes the performance of computational runtime (ms) of stock trend feature extraction using the proposed approach on large datasets. From the Table 3, it is clearly shown that the present feature extraction procedure has low computation runtime as compared to the existing approaches.

Table. 3 Stock trend features extraction using the proposed

MI	Chi-Square	Rough set	IG	GA	PSO	Proposed FS
64	63	66	60	56	60	48
64	52	77	62	59	66	47
57	58	80	56	56	53	48
62	50	72	64	58	66	38
55	63	65	51	58	65	35
60	53	77	54	54	60	47
60	68	71	54	63	72	46
63	73	66	64	51	68	40
53	54	78	54	58	52	46
53	74	71	57	52	52	40
61	73	80	61	62	59	46
58	55	75	56	57	58	48
57	56	64	54	58	51	44
58	70	74	60	59	72	42
64	60	81	50	65	61	42
55	66	65	55	52	69	36
61	69	60	59	65	50	48
52	62	84	56	53	58	47
54	51	64	63	53	72	44
64	57	66	50	50	63	37

Table. 4 Performance Analysis Of Computational Runtime (Ms) With Different Traditional Feature Selection Models

Features Size	**MI**	**Chi-square**	**Roughset**	**Information Gain**	**Genetic Algorithm**	**PSO**	**Proposed FS**
StockID-100	5417	6297	7230	7024	6638	6481	4747
StockID-200	5365	6529	6957	5917	6253	6516	3965
StockID-300	6200	6609	5959	6882	7099	6469	3474
StockID-400	6122	5514	6055	5729	7430	5584	3495
StockID-500	6727	5676	7488	6685	6103	6753	4809
StockID-600	5774	6089	6823	6273	5996	6597	4292
StockID-700	7205	6045	6060	7449	7448	6319	3888
StockID-800	6976	7340	6036	6544	6855	5864	4723
StockID-900	7080	5710	6692	5661	6866	7365	4488
StockID-1000	5854	7118	7087	5916	5541	5552	4368
StockID-1100	6431	7224	5597	7359	7022	6253	4470
StockID-1200	6595	5767	5989	7086	5950	7552	3995
StockID-1300	6077	7018	5357	6242	5645	6886	4335
StockID-1400	7350	6953	5711	6927	7133	7061	4439
StockID-1500	7536	5695	5414	6211	6909	6438	4303
StockID-1600	7557	7425	6421	6729	6922	5686	3618
StockID-1700	7032	5520	6334	5614	5942	5840	4524
StockID-1800	7556	7183	5417	5700	7236	5630	4734
StockID-1900	5669	5589	5708	7281	5900	6588	4568
StockID-2000	6318	6253	5603	6852	5925	7090	4178

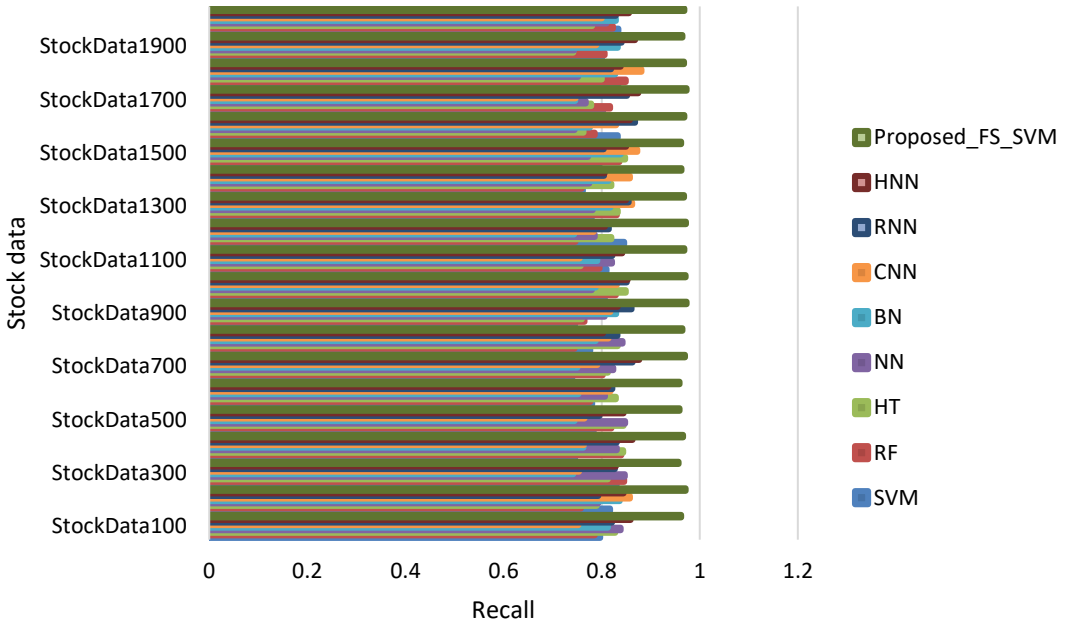

Figure. 11 Performance analysis of recall using different traditional feature selection based classification frameworks

4.1 Recall

Fig. 11 describes the performance of recall of stock trend classification using the proposed learning framework on large datasets. From the Fig. 11 it is clearly shown that the present framework has high computational recall as compared to the existing frameworks.

Fig. 12 describes the performance of precision of stock trend classification using the proposed learning framework on large datasets. From the Fig. 12 it is clearly shown that the present framework has high computational precision as compared to the existing models.

4.2 Precision

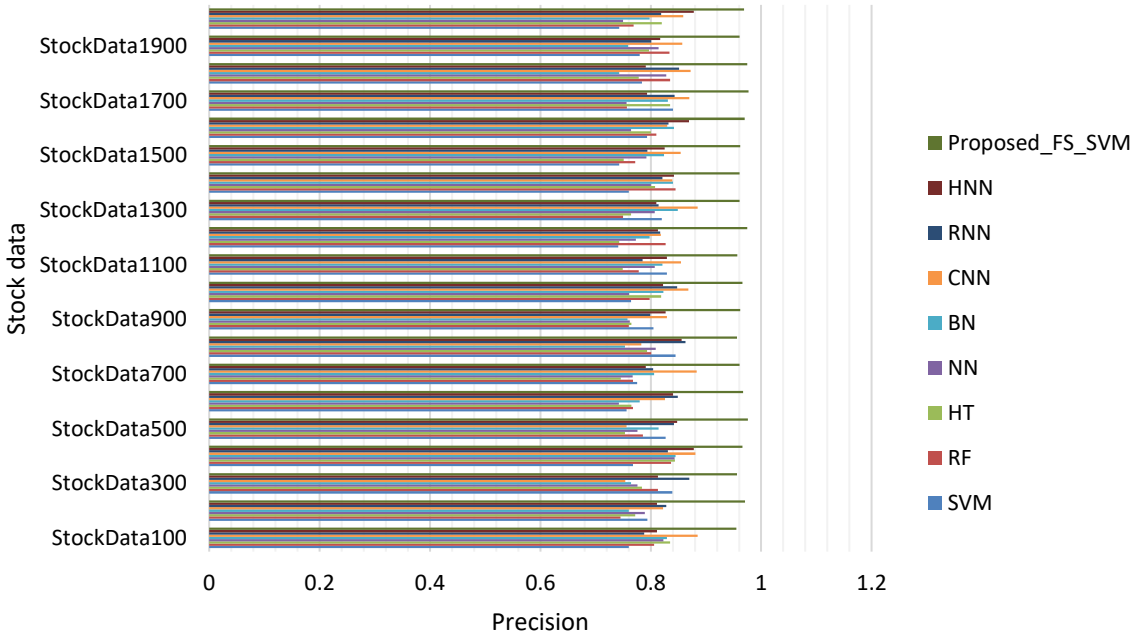

Figure. 12 Performance analysis of Precision using different traditional classification learning frameworks

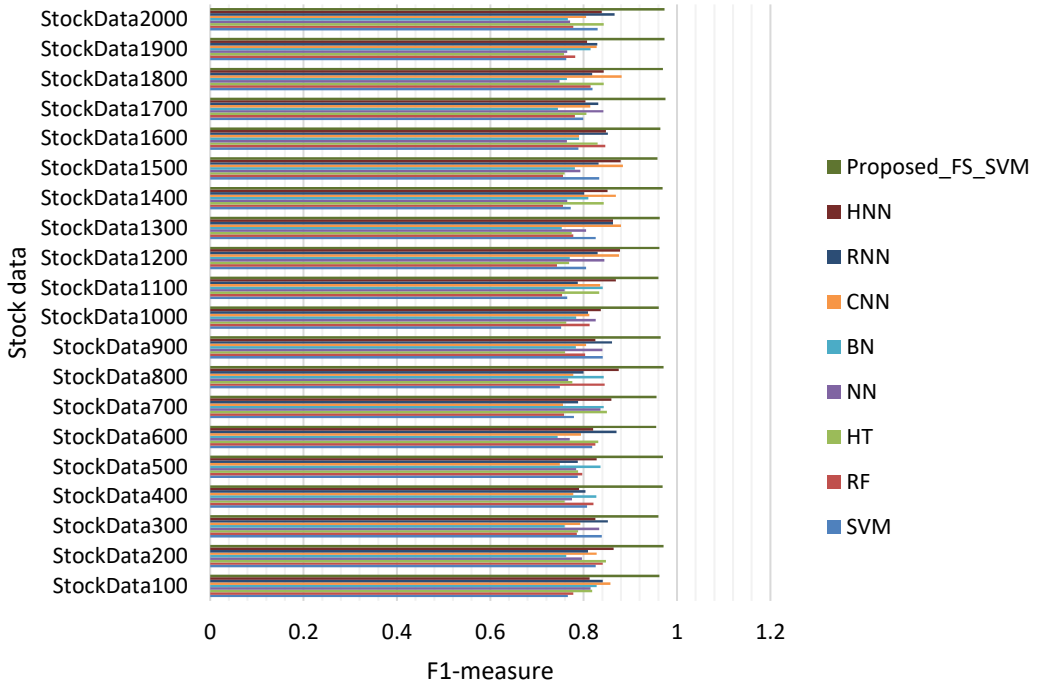

Figure. 13 Performance analysis of F1-Measure using different traditional classification learning frameworks

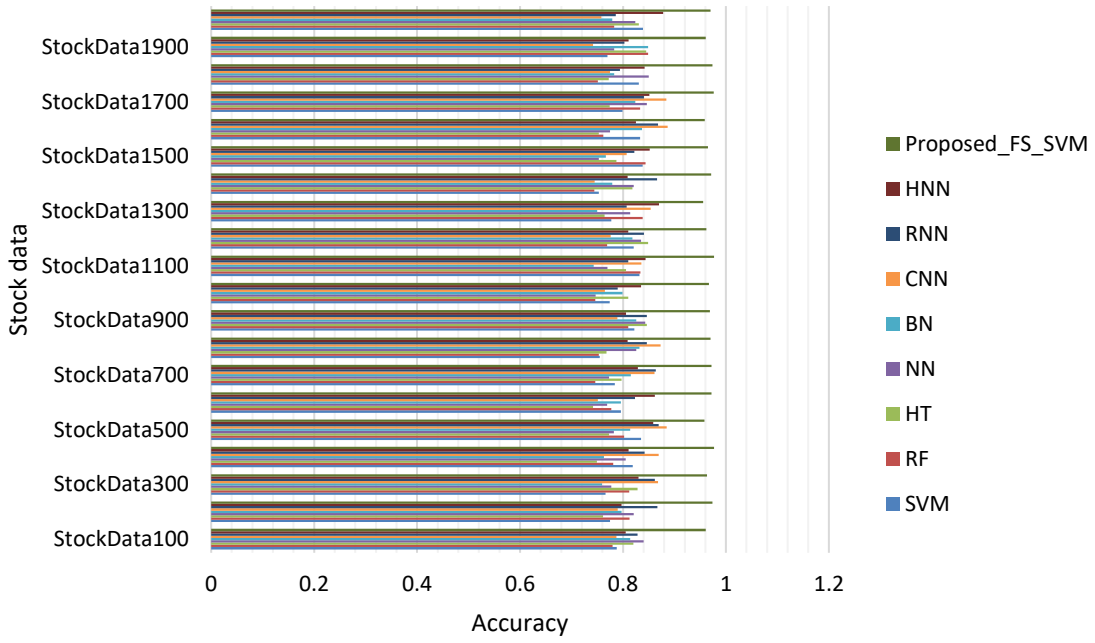

Figure. 14 Performance analysis of Accuracy using different traditional classification learning frameworks

4.3 F1-measure

Fig. 13 describes the performance of F1-measure of stock trend classification using the proposed deep learning framework on large datasets. From the Fig. 13 it is clearly shown that the present framework has high computational F1-measure as compared to the existing frameworks.

4.3 Accuracy

Fig. 14 describes the performance of accuracy of stock trend classification using the proposed deep learning framework on large datasets. From the Fig. 14 it is clearly shown that the present framework has high computational accuracy as compared to the existing frameworks.

5. Inference

As described in the above results, the data filtering and feature selection of the proposed model improves the performance of the various measures such as precision, recall, F-measure, accuracy and runtime. From the above tables, it is clearly identified that the proposed feature extraction and scoring approach optimizes the stock sentiment of the social media comments and its technical data. As compared to the traditional feature extraction measures, proposed stock feature extraction function has high computational efficiency with less runtime in the real-time stock market databases.

In the experimental results, nifty 50 stocks and its technical data are taken as input data to the proposed model. Initially, input data is pre-processed using the data transformation approach. Here, the filtered data is used as training data to the classification model. In the above tables, various feature extraction measures such as mutual information (MI), chi-square, roughset, information gain and genetic algorithm are used to find the essential features on the input stock data for classification problem. The output of the classification approach is stock trend prediction i.e buy(1) or sell(0).

From the results, it is observed that the proposed model has 7% efficiency for features identification and runtime (ms).

The performance of the proposed non-linear SVM classifier is compared to the traditional classifiers such as SVM, random forest (RF), hoeffding tree (HT), neural network (NN) and Bayesian networks. From the above tables, it is observed that the performance of the proposed non-linear classifier is better than the traditional classifiers in terms of recall, precision, F-measure, accuracy and runtime (ms). Also, approximately on an average 8-10% accuracy is optimized in the proposed model than the traditional stock market prediction classifiers. Discussions: Table 2-4 describes the efficiency of the hybrid feature selection approach to the conventional feature selection models on stock market data. From the results, it is noted that the proposed approach has better efficiency in selecting the features for classification problem. Section 4.1-4.4 illustrates the performance of classification measures on the real-time stock market data.

6. Conclusion

In this paper, a new sentiment and technical based stock market trend prediction model is designed and implemented on real-time market data. Most of the existing technical indicators are difficult to predict the bullish or bearish trend by using the technical data and social stock comments. Also, these indicators contain noise during data pre-processing and stock feature extraction. In the proposed work, a new technical indicator to the stock market data is proposed to find the bullish or bearish trend in each stock. Here, social media stock related comments are used to find the movement of the stock or trend of the stock along with the technical indicator. Experimental results proved that the present model has high computational efficiency than the traditional technical indicators in terms of accuracy, f-measure, precision and recall. From the experimental results, it is observed that the proposed stock market trend prediction model has 7% of runtime (ms) and 10% of average classification accuracy as compared to the traditional trend prediction models on training and test dataset.

Conflicts of Interest

Authors declare that he has no conflict of Interest.

Author Contributions

The entire work of conceptualization, formal analysis, validation, and writing, editing and modification of article were done by Lakshmana Phaneendra Maguluri under the supervision of Ragupathy Rengaswamy.

References

[1] K. Gadiaa and K. Bhowmick, “Parallel Text Mining in Multicore Systems Using FP-tree Algorithm”, *Procedia Computer Science,* Vol. 45, pp. 111-117, 2015.

[2] J. Bollen, H. Mao, and X. Zeng, “Twitter mood predicts the stock market”, *Journal of Computational Science*, Vol. 2, No. 1, pp. 1-8, March 2011.

[3] F. Xianghua, L. Wangwang, X. Yingying, and C. Laizhong, “Combine How Net lexicon to train phrase recursive auto encoder for sentence-level sentiment analysis”, *Neuro Computing*, Vol. 241, pp. 18-27, 2017.

[4] D. Lin, L. Li, D. Cao, Y. Lv, and X. Ke, “Multi-modality weakly labeled sentiment learning based on Explicit Emotion Signal for Chinese micro blog”, *Neurocomputing*, Vol. 272, pp. 258-269, 2018.

[5] M. Giatsoglou, M. G. Vozalis, K. Diamantaras, A. Vakali, G. Sarigiannidis, and K. C. Chatzisavvas, “Sentiment analysis leveraging

emotions and word embeddings", *Expert Systems with Applications*, Vol. 69, pp. 214-224, 2017.

[6] K. Guo, Y. Sun, and X. Qian, "Can investor sentiment be used to predict the stock price? Dynamic analysis based on China stock market", *Physica A: Statistical Mechanics and its Applications*, Vol. 469, pp. 390-396, 2017.

[7] C. Hung, "Word of mouth quality classification based on contextual sentiment lexicons", *Information Processing & Management*, Vol. 53, No. 4, pp. 751-763, 2017.

[8] O. Araque, I. C. Platas, J. F. S. Rada, and C. A. Iglesias, "Enhancing deep learning sentiment analysis with ensemble techniques in social applications", *Expert Systems with Applications*, Vol. 77, pp. 236-246, 2017.

[9] D. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)", In: *Proc. of Learning Representations*, 2016.

[10] T. Chen, X. He, and MY. Kan, "Context-aware image tweet modeling and recommendation", In: *Proc. of 24th ACM international conference on Multimedia*, pp. 1018-1027, 2016.

[11] X. He, H. Zhang, M. Y. Kan, and T. S. Chua, "Fast matrix factorization for online recommendation with implicit feedback", In: *Proc. of ACM SIGIR conference on Research and Development in Information Retrieval*, pp. 549-558, 2016.

[12] L. P. Maguluri and R. Ragupathy, "A New sentiment score based improved Bayesian networks for real-time intraday stock trend classification", *International Journal of Advanced Trends in Computer Science and Engineering*, Vol. 8, No. 4, 2019.

[13] P. C. Chang, C. H. Liu, J. L. Lin, C. Y. Fan, and C. S. P. Ng, "A neural network with a case based dynamic window for stock trading prediction", *Expert Systems with Applications*, Vol. 36, No. 3, pp. 6889-6898, 2009.

[14] B. B. Nair, V.P. Mohandas, and N. R. Sakthivel, "A Decision Tree- Rough Set Hybrid System for Stock Market Trend Prediction", *International Journal of Computer Applications*, Vol. 6, No. 9, pp. 1-6, 2010.

[15] B. B. Nair, V. P. Mohandas, and N. R. Sakthivel, "A Stock Market Trend Prediction System Using a Hybrid Decision Tree-Neuro-Fuzzy System", In: *Proc. of Advances in Recent Technologies in Communication and Computing*, pp. 381-385, 2010.

[16] R. Majhi, G. Panda, and G. Sahoo, "Development and performance evaluation of FLANN based model for forecasting of stock markets", *Expert Systems with Applications*, Vol. 36, No. 3, pp. 6800-6808, 2009.

[17] A. P. Ratto, S. Merello, L. Oneto, Y. Ma, L. Malandri, and E. Cambria, "Ensemble of Technical Analysis and Machine Learning for Market Trend Prediction", In: *Proc. of IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 2091-2096, 2018.

[18] T. Loughran and B. Mcdonald, "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks", *The Journal of finance*, pp. 35-65, 2011.

[19] R. Blagus and L. Lusa, "Class prediction for high-dimensional class-imbalanced data", *BMC Bioinformatics*, pp. 1-17, 2010.

[20] R. Luss and A. d'Aspremont, "Predicting abnormal returns from news using text classification", *Journal Quantitative Finance*, Vol. 15, No. 6, pp.1-14, 2015.

[21] Y. Ma, H. Peng, and E. Cambria, "Targeted Aspect-Based Sentiment Analysis via Embedding Commonsense Knowledge into an Attentive LSTM", In: *Proc. of Artificial Intelligence Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 5876-5883, 2018.

[22] S. Merello, A. P. Ratto, L. Oneto, and E. Cambria, "Predicting Future Market Trends: Which Is the Optimal Window?", *Recent Advances in Big Data and Deep Learning*, pp. 180-185, 2019.

[23] A. Mittal and A. Goel, "Stock Prediction Using Twitter Sentiment Analysis", *Stanford University*, CS229, pp. 1-5, 2011.

[24] P. Areekul, T. Senjyu, H. Toyama, and A. Yona, "Notice of Violation of IEEE Publication Principles: A Hybrid ARIMA and Neural Network Model for Short-Term Price Forecasting in Deregulated Market", *IEEE Transactions on Power Systems*, Vol. 25, No. 1, pp. 524-530, 2009.

[25] G. E. P. Box, G. Jenkins, G. Reinsel, and G. Ljung, "Time series analysis: Forecasting and control", *Journal of Time*, Vol. 31, pp. 238–242, 1976.

[26] S. K. Chandar, M. Sumathi, and S. N. Sivanandam, "Prediction of Stock Market Price using Hybrid of Wavelet Transform and Artificial Neural Network", *Indian Journal of Science and Technology*, Vol. 9, No. 8, pp. 1-5, 2016.

[27] M. Syamala and N. J. Nalini, "A Filter Based Improved Decision Tree Sentiment Classification Model for Real-Time Amazon Product Review Data", *International Journal of*

Intelligent Engineering and Systems, Vol. 13, No. 1, pp. 191-202, 2020.

[28] Y. Zhai, W. Song, X. Liu, L. Liu, and X. Zhao, "A Chi-Square Statistics Based Feature Selection Method in Text Classification", In: *Proc. of Software Engineering and Service Science (ICSESS)*, Beijing, China, pp. 160-163, 2018.

[29] X. Wang, B. Guo, Y. Shen, C. Zhou, and X. Duan, "Input Feature Selection Method Based on Feature Set Equivalence and Mutual Information Gain Maximization", *IEEE Access*, Vol. 7, pp. 151525-151538, 2019.

[30] Y. Wang, "Unsupervised Representative Feature Selection Algorithm Based on Information Entropy and Relevance Analysis", *IEEE Access*, Vol. 6, pp. 45317-45324, 2018.

[31] H. Zhao, P. Wang, Q. Hu, and P. Zhu, "Fuzzy Rough Set Based Feature Selection for Large-Scale Hierarchical Classification", *IEEE Transactions on Fuzzy Systems*, Vol. 27, No. 10, pp. 1891-1903, 2019.

[32] K. Nag and N. R. Pal, "A Multiobjective Genetic Programming-Based Ensemble for Simultaneous Feature Selection and Classification", *IEEE Transactions on Cybernetics*, Vol. 46, No. 2, pp. 499-510, 2016.

[33] B. Tran, B. Xue, and M. Zhang, "Variable-Length Particle Swarm Optimization for Feature Selection on High-Dimensional Classification", *IEEE Transactions on Evolutionary Computation*, Vol. 23, No. 3, pp. 473-487, 2019.