TEST Engineering & Management

January - February 2020
ISSN: 0193 - 4120 Page No. 64 - 72

A Hybrid GA-SVM and Sentiment Analysis for Forecasting Stock Market Movement Direction

Shashi Shekhar, Neeraj Varshney
Department of Computer Engineering & Applications
GLA, University, Mathura, India-281406
shashi.shekhar@gla.ac.in ,neeraj.varshney@gla.ac.in,

Article Info
Volume 82
Page Number: 64 - 72
Publication Issue:
January-February 2020

Abstract
For a company, important indicator is stock price. This value is affected by various factors. Differently, public emotion and sentiment can be affected by various events. Price of stock market can be affected by this. Price of stock are not a static one due to dependency of various factors. They are nonlinear time series data, has high amount of noise and dynamic. To this area of research, machine learning algorithms may be applied to solve problems in nonlinear predictions because of its high learning ability.

For predicting stock price, learning based methods are used and its performance can be enhanced by various methods. Still it is challenging task to predict stock market. Psychology of investors are reflected by User-generated textual content provided by internet. On stock market, important role is played by sentiment of investor and it is used for prediction of stock price. Support vector machine based on Genetic algorithm is integrated with machine learning based sentiment analysis.

An improvement about 18.6% in accuracy is obtained by combing sentiment variable. Final accuracy is about 89.93%. Risk of investors can be reduced and they are allowed to make wide range of decisions by combining proposed method with stop-loss order strategy. Asset fundamental value information is contained by sentiment. Stock market can be indicated by this in an effective way. Time interval can be expanded in future for gathering huge amount of textual documents.

Keywords: *Genetic Algorithm (GA), Text Mining, Stock Markets, Sentiment Analysis, Decision Making, support vector machine (SVM), Day-Of-Week Effect.*

Article History
Article Received: *14 March 2019*
Revised: *27 May 2019*
Accepted: *16 October 2019*
Publication: *01 January 2020*

1 Introduction

In finance, most important issue is forecasting price of stock. Investors are highly concentrate on this. Accurate stock price forecasting are has rewards that are very attractive and gives profit. Irreparable consequences are caused by unreliable and inaccurate predictions. So, stock prices should be predicted by an efficient model.

In the world, most complex signal corresponds to financial time series and it is highly noisy one. Prediction of price of stock is a highly difficult task [1]. In Random Walk Theory (RWT) andEfficient Market Hypothesis (EMH), this makes belief of various economists [2]. Random walk is followed by financial market and it is unpredictable. Advancements in technology made it as a predictable one, but better results are not produced by existing methods.

In this area of research, prediction of stock market is a challenging task. In the world, it is most popular topic of research [3]. Sentiment of investors are reflected by a sourced provided by internet which includes, textual content, social networking

January - February 2020
ISSN: 0193 - 4120 Page No. 64 - 72

websites, forums, blogs, news. Price of stock can be predicted as a time series of data.

From huge amount of textual documents, knowledge can be extracted by automated approach [4]. From opinionated contents, emotions, attitudes and views can be extracted automatically by sentiment analysis [5]. Sentiment indexes are constructed using sentiment analysis. Movement direction is forecasted by aggregating those indexes with data of stock market. Consider the effect of day-of-week in order to obtain sentiment index in an effective manner. On weekends and weekdays, huge amount of information may be generated.

Computation of sentiment index has sever effect of day-of-week. Stock market is dynamic, evolutionary and nonlinear. Because of this properties, it is more difficult to predict the movement direction of stock. Nonlinear problem can be solved by using Support vector machine (SVM). It provides solution which unique as well as globally optimum. In feature space, maximal margin hyper-plane is selected to reduce the problem of overfitting.

Fivefold cross validation is implemented to address this issue. But it leads to bias. Bias is eliminated by integrating realistic rolling window method with SVM based on genetic algorithm. In stock market movement direction forecasting, better performance is exhibited by combing stock market data with sentiment features. It reduces the role of sentiment of investor on stock market. Asset fundamental value information is contained by sentiment. Stock market can be indicated by this in an effective way. Time interval can be expanded in future for gathering huge amount of textual documents.

Organization of the paper is as follows. Related works are reviewed in section II. SVM and new sentiment analysis techniques are described elaborately in section III. Results are presented in section IV. Conclusion and future enhancements are presented in section V.

2 Related work

For time series data prediction, time series models are compared with ANN by various researchers. In [6], prices of Istanbul Stock Exhange (ISE) National 100 Index is predicted using support vector machines (SVM) and ANN. In input, used ten technical indicators to obtain maximum of 75.74% accurate results by SVM and ANN with kernel of polynomial. In time series data analysis, test set should be advanced than training set.

NASDAQ index's rate of exchange are predicted by using different ANN models in [7]. When compared with other model, better performance is shown by classical ANN. Worst performance is exhibited by GARCH model based on dynamic architecture of ANN (DAN2). On Tehran Stock Exchange, prices of stock are predicted by using ANN in [8]. Principal component analysis (PCA) method is used to select efficient factors. To Japanese stock market, ANN is applied in [9].

Simulated annealing methods and genetic algorithms are used to enhance ANN's accuracy in prediction and rectifies, back propagation algorithm's problem of local convergence.

SVM is trained by selecting features from input and it has 19 technical indicators to choose. Fractal selection method is used for selecting features. With radius basis kernel functions, SVM are trained. In Shanghai Stock Exchange Composite Index (SSECI) trend prediction, better performance is shown by SVMs with fractal feature selection. Four feature selection methods are combined with SVM in [11]. When compared to individual SVM, better performance is shown by four hybrid classifiers.

Wu et al. [12] integrated SVM with sentiment analysis. Autoregressive conditional heteroskedasticity is generalized for exploring relation among stock forum sentiment and volatility of stock price. In volatility terms, financial risks are measured efficiently by this method.

January - February 2020
ISSN: 0193 - 4120 Page No. 64 - 72

At the point when these calculations are activated to make exchanges, the market floods or falls radically in light of the high unpredictability made by the calculations in the market [13]. Taking into account that the high-recurrence exchanging represents more than 70 percent of dollar exchanging volumes the U.S. budgetary market, and those blaze crashes occurred more than multiple times somewhere in the range of 2006 and 2011 [14], guaging the glimmer crash and having the option to forestall any misfortune are unequivocally required for the unarmored normal individual financial specialists' security, solid market environment, and the entire economy.

There are various methods for predicting direction of stock market movement which includes neural network and deep learning methods.

It uses many learning, relapse, characterization, neural systems calculations, for example, bolster vector machine, arbitrary timberland, strategic relapse, credulous Bayes, and repetitive neural systems, and attempts to make precise expectations by modifying itself as per the market changes [15]. Another prominent strategy is to utilize characteristic language handling procedures that let machines extricate and comprehend data composed and communicated in human dialects, and attempt to catch financial exchange assessments for settling on speculation choices dependent on the disposition or the notions of the securities exchange [16].

Stock market is forecasted by traditional and modern financial engineering. It used technical and fundamental analysis. Stocks intrinsic values are computed using fundamental analysis. It is done based on economic status and performance of company. Volume and price dynamics are focused by technical anlysis. It develops technical indicators for capturing investment time[17].

Stock market is studied using theories of network science. In order to compute major influencer, stock market's properties of network structure are analysed. Stock market communities are also detected by this [18]. Network science is use dby few methods to forecast stock market's movement direction.

One of only a handful hardly any investigations constructed corporate news systems utilizing top 50 European organizations in STOXX 50 file as hubs, and the whole of the quantity of news things with the basic subject of each organization match as connection loads. This examination discovered that the normal eigenvector centrality of the news systems affects return and instability of the STOXX 50 file [19].

By arranging the hubs into three sorts (Hub, Periphery, and Connector) as per the hub's connectedness, this investigation made 9 diverse connection types, and found that the time-arrangement of portion of the connection type P-H and C-H have a prescient power with the most extreme exactness of 69.2% [20].This paper not just joins SVM with a moving window way to deal with make our strategy increasingly significant and pragmatic in a money related area, yet in addition incorporates supposition examination into an AI technique dependent on SVM to conjecture securities exchange development bearing with thought of human feelings.

3 Proposed Methodology

Role of investors are emphasized to forecast the direction of stock market in this work. Stock market is driven by psychology of investor and it reflected by content generated by user in internet. It is a primary source. Daily sentimental indexes are formed by converting unstructured textual documents in sentiment analysis.

Irregularity in financial conditions of days of a week effects this. It means return filled on Monday will be very less when compared to other days. Sentiment index precision is affected by this. On weekdays, past sentiment changes are introduced with exponential function and they are generalized to get a value on holidays. For experimentation, financial

January - February 2020
ISSN: 0193 - 4120 Page No. 64 - 72

websites like, Eastmoney and Sina Finance are chosen. Financial review data's corpus is obtained by this.

The SSE 50 index is predicted using machine learning model GA-SVM. In China, it is an significant index and it is computed by the implementation of realistic rolling window method and fivefold cross validation. In movement direction forecasting, better results can be achieved by combing stock market data with sentiment features. Figure 1 shows stock market prediction architecture overview.

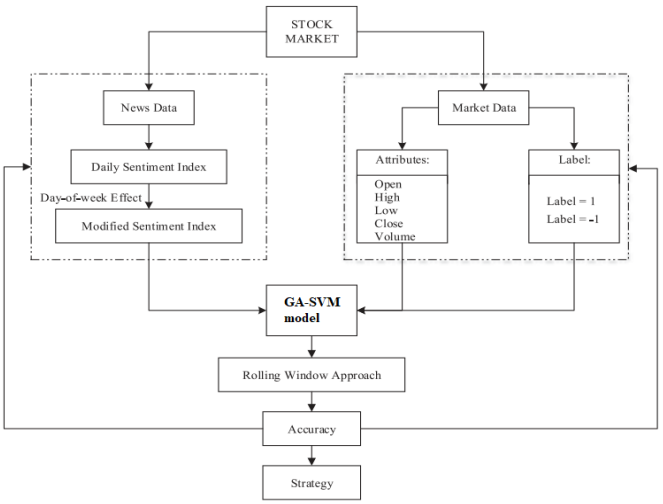

Fig. 1. Overview of stock market prediction architecture

Investor Sentiment: There are three steps in this section. From internet, target textual documents are downloaded automatically by constructing web crawler at first and based on corpus, daily sentiment indexes are constructed. Finally, on day-of-week effect, adjustments are considered.

Step 1 Web crawler:From internet, target textual documents are downloaded automatically by constructing web crawler. For future processing, they are stored in database. Figure 2 shows this framework. Seeds are used start web crawler. URL list corresponds to seeds. URL queue is managed by scheduler. Priority is decided by this and duplicated parts are eliminated.

From internet, web pages are acquired by downloader. Spider are given with those pages. Pages are parsed by this. Target contents are extracted. Two sections need to be obtained. They are, from websites, textual news including date and in database they are stored. Another one is page URL and scheduler s formed by transposing URL.

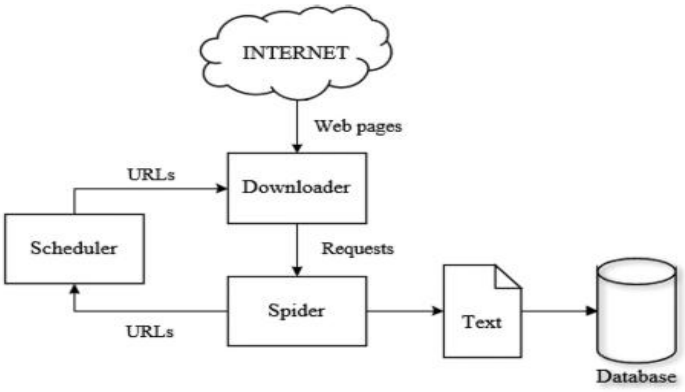

Fig. 2. Framework of the web crawler**.**

Until obtaining stable target textual documents, repeat this procedure. In database, using contents, headline and time, display every document.

Step 2 Daily sentiment:

During specific period of time, textual data are processed using sentence-based sentiment analysis method. Instead of one word, whole document is interpreted by sentence. Complete meaning can be expressed by sentence. Problem of ambiguity can be addressed by this.

First, sentences are formed by dividing documents. Separate words are formed by segmenting those sentences. On sentiment space, those word are projected and number of negative and positive words are counted. Based on Chinese Sentiment Analysis Ontology Base and HowNet, sentiment values are assigned and every sentence's polarity is decided.

Online common-sense knowledge base is HowNet. Concepts inter attribute and inter conceptual relationships are unveiled by this. In Chinese and their English equivalents lexicons, they are connoted. Dalian University of Technology constructed Chinese Sentiment Analysis Ontology Base. Various aspects having part of sentiment

TEST Engineering & Management

January - February 2020
ISSN: 0193 - 4120 Page No. 64 - 72

intensity, polarity and speech are used to represent phrases and words. Every document is classified after that. Compute daily sentiment index $S_t$as,

$$S_t = \begin{cases} \dfrac{2M_t^{bull}}{\left(M_t^{bull} + M_t^{bear}\right) - 1} & M_t^{bull} > M_t^{bear} \\ 0 & M_t^{bull} = M_t^{bear} \\ 1 - \dfrac{2M_t^{bear}}{\left(M_t^{bull} + M_t^{bear}\right)} & M_t^{bull} < M_t^{bear} \end{cases}$$

Where, number of positive comments is represented by M_t^{bull}, number of negative comments is represented by M_t^{bear} in day t. Value of $S_t$lies between –1 to 1, where, neutral position of people is represented by 0 value. Positive view of people is represented by greater than 0 value and negative view is represented by less than 0 value

Step 3 Modified sentiment:

Commonly known financial anomalies is effect of day-of-week [5]. When compared to other days, Monday there will be average return. This is because, on week end large number of news will be reported.

On Monday, mind of investors may be changed in tasking decision based on valuable information available to them. In order to boost public image and for ensuring stock's stability, on weekend, important news are released by corporations. Investors are going to have more time to accept, if they receive bad news and they will look into process of spreading the news to more people, if they receive good news.

Exponential time function is used for gauging effect to Monday from Saturday. More recent news are having high impact. On stock market, on changes of past price, using an exponential function, sentiment measure is incorporated. Past sentiment's weighted average defines Monday sentiment. Weights are decreased exponentially and it is expressed as,

$$S_m = e^{-\lambda t_1} S_1 + e^{-\lambda t_2} S_2 + e^{-\lambda t_3} S_3$$

Where, Saturday sentiment is represented by S_1, Sunday sentiment is represented by S_2 and Monday sentiment is represented by S_3. Modified Monday sentiment is given by Sm. Prescribe $\lambda(\lambda > 0)$. t1 = 2, t2 = 1, and t3 = 0. On national holidays as well as some special days, stock market will be closed. So, for common situations, Sm is generalized. On stock market, if we have n holiday days, on n + 1th day, sentiment is represented as,

$$S_{n+1} = e^{-n\lambda} S_1 + e^{-(n-1)\lambda} S_2 + \cdots + e^{-\lambda} S_n + S_{n+1}.$$

Optimization of the SVM Classifier by Genetic Algorithm (GA): SVM performance is depends on the selection of parameters of SVM. Following parameters has to selected in RBF-SVM.

1) Regularization parameter C:Trade off between model complexity and SVM model's fitting error are computed by this. It is also termed as SVM's cost parameter.
2) Kernel free parameterσ : RBF kernel bandwidth is computed by this. Mapping of input space to feature space with high dimension is defined by this. Optimum value of C and σ are computed by GA. Ability in generalization and accuracy of prediction are ensured by this.

Principles of GA:

Natural selection process is modelled using adaptive optimization technique known as GA. In machine learning and optimization algorithms, it is applied successfully. For an optimization problem, potential solutions are given by chromosome population in GA. After iterative process, obtain optimum solution.

January - February 2020
ISSN: 0193 - 4120 Page No. 64 - 72

Randomly generate initial population. Optimum value of C and σ parameters are computed by using GA. There are two parts in every chromosome $X = [b_C, b_\sigma]$. As illustrated in figure 3, binary coding system is used to construct it. Generational evolution method is implemented by GA after initialization and for searching, following genetic operations are applied to compute optimum solution.

(1) Fitness Evaluation: In every iteration, performance of every chromosome is assessed by designing fitness function. Maximization of SVM classifier's accuracy in predicting faults in gearbox is the major goal of this search technique. value of fitness function is maximized by designing GA to fit C and σ values.

(2) Parent Selection: For reproduction, proper parents are selected in this step. Population is added with experience by adapting heuristic method with a generation gap. Selection individual is determined by this generation gap. 0.9 is used as a generation gap in this work. For following mutation and crossover operation, 90% of population with better value of fitness are maintained.

(3) Crossover: In problem space, which has to explored, new solution regions are allowed by crossover operation. Cross over a selected chromosome in one or more positions assigned randomly in crossover operation. Off spring are generated by combining rest of chromosome with resulted chromosome. In next generation, new population are formed by replacing old population with generated offspring. Figure 4 illustrates the crossover operation process. Probability of crossover is 0.7.

(4) Mutation: 1" and "0" are mutated with each other, during mutation in binary coding system. With a given probability of 0.7, mutate every chromosome element. Figure 5 illustrates the process of mutation.

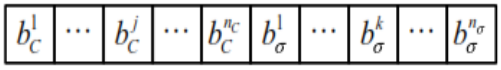

Fig. 3. Encoding in a chromosome, where n_C and n_σ are the numbers of binaries for C and σ respectively.

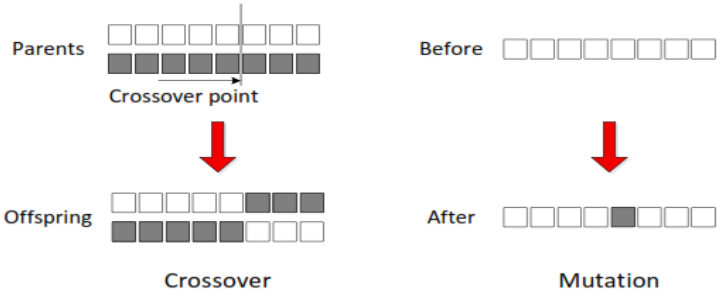

Fig. 4. The crossover and mutation operations.

Proposed GA-SVM Hybrid Classifier:

Multiclass fault in drivetrain gearboxes are identified by proposed GA-SVM hybrid classifier. Optimum values of C and σ are computed using GA. SVM's ability in generalization and prediction accuracy are ensured by this optimum values. Figure 7 shows the flowchart of proposed classifier. Selected databases are fitted with SVM by computing global optimum solutions by GA. Identification of gearbox fault is done by SVM classifier with these optimum parameters. There are certain limit for C and σ values in order to ensure SVM's generalization capability.

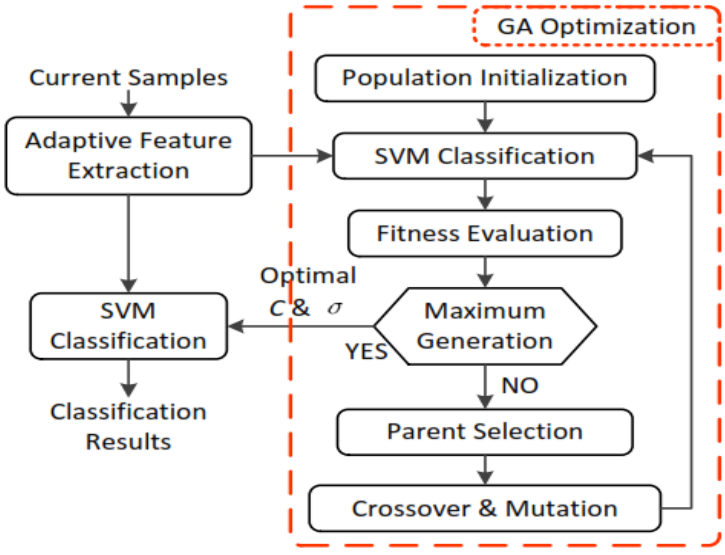

Fig. 5. The proposed GA-SVM hybrid classifier.

TEST Engineering & Management

January - February 2020
ISSN: 0193 - 4120 Page No. 64 - 72

4 Experimental Results and Discussions

Index SSE 50's trend is explored by this method. It is a significant index in China. It used data of stock market, news documents relating to it with its constituents. On Shanghai stock market, it is a primary blue-chip stock index. Good liquidity's 50 largest stock and its representatives are used to made this.

Percentage change, RMB change, RMB's trading volume, number share's trading volume, low of day, high of day, price at closing and price at opening are the time series data. From Wind Economic Database, these data of SSE 5 index are downloaded with its constituents. In China's financial information service industry market, Wind Economic Database are the leading provider.

Over the period of 17.06.2014 to 7.06.2016, from Eastmoney stock and Sina stock forum, all post of 51 shares are downloaded which includes 485 days of trading. Every stock has 37,855 reviews and peak value of 23,236 and lowest value of 7,797. Table I shows these details. There are 19,30,592 reviews on both Eastmoney stock and Sina stock forum.

4.1 Running Time Comparison Results

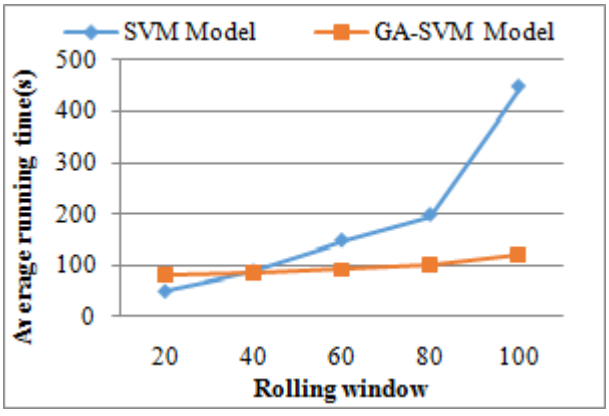

Fig.6. Running Time comparison Results

Figure 6 shows the runtime comparison of SVM and GA-SVM methods. Increase in image size increase time linearly. When compared with SVM, proposed method has less running time as shown by figure 6. With huge amount of data, proposed method can produce better results in an effective manner.

4.2 Detection Accuracy

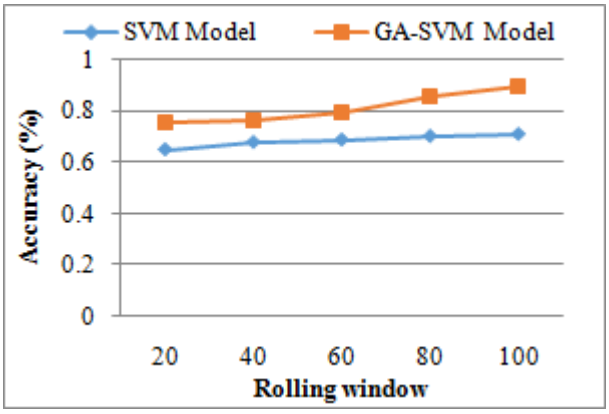

Fig.7. Detection Accuracy comparison Results

Accuracy metric comparison of SVM and GA-SVM are shown in figure 8. On different test images with various size, high accuracy value is produced by proposed GA-SVM model. Increase in test set size, increases accuracy value of GA-SVM. Few test sets may produces less value of accuracy. This is caused by difference in training and test set. Stock market's more information about its features are captured by GA-SVM. Performance of prediction is enhanced by including sentiment feature with baseline model. Due to inclusion of investor's sentiment, accuracy value is enhanced to 89.93%

4.3 Recall Comparison Results

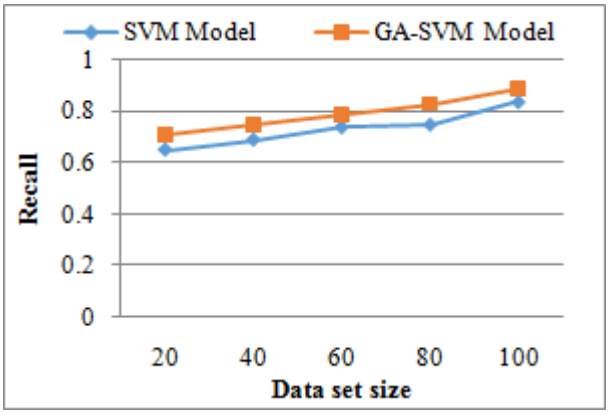

Fig.8. Recall comparison Results

Recall metric comparison of SVM and GA-SVM are shown in figure 8. On different test images with various size, high recall value is produced by

January - February 2020
ISSN: 0193 - 4120 Page No. 64 - 72

proposed GA-SVM model. Increase in test set size, increased recall value of GA-SVM. Few test sets may produces less value of recall. This is caused by difference in training and test set. When compared with SVM, high rate of convergence and recall value are produced by GA-SVM.

5 Conclusion and Future work

Role of investors are emphasized to forecast the direction of stock market in this work. Stock market is driven by psychology of investor and it reflected by content generated by user in internet. It is a primary source. Daily sentimental indexes are formed by converting unstructured textual documents in sentiment analysis.

Irregularity in financial conditions of days of a week effects this. It means return filled on Monday will be very less when compared to other days. Sentiment index precision is affected by this. On weekdays, past sentiment changes are introduced with exponential function and they are generalized to get a value on holidays. For experimentation, financial websites like, East money and Sina Finance are chosen. Financial review data's corpus is obtained by this.

The SSE 50 index is predicted using machine learning model GA-SVM. In China, it is an significant index and it is computed by the implementation of realistic rolling window method and fivefold cross validation. In movement direction forecasting, better results can be achieved by combing stock market data with sentiment features.

An improvement about 18.6% in accuracy is obtained by combing sentiment variable. Final accuracy is about 89.93%. Risk of investors can be reduced and they are allowed to make wide range of decisions by combining proposed method with stop-loss order strategy. Asset fundamental value information is contained by sentiment. Stock market can be indicated by this in an effective way. Time interval can be expanded in future for gathering huge amount of textual documents.

References

1. S. D. Patel, D. Quadros, V. Patil, M. Pawale, and HarshaSaxena, "Stock prediction using neural networks," Int. J. Eng. Manag. Res., vol. 7, no. 2, pp. 490–493, 2017.
2. J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," J. Comput. Sci., vol. 2, no. 1, pp. 1–8, 2011.
3. I. Perikos and I. Hatzilygeroudis, "Recognizing emotions in text using ensemble of classifiers," Eng. Appl. Artif. Intell., vol. 51, pp. 191–201, 2016.
4. B. Wu, X. Zhou, Q. Jin, F. Lin, and H. Leung, "Analyzing social roles based on a hierarchical model and data mining for collective decisionmaking support," IEEE Syst. J., vol. 11, no. 1, pp. 356–365, Mar. 2017.
5. B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis," Mining Text Data. New York, NY, USA: Springer, 2012.
6. Kara Y, Boyacioglu MA, Baykan ÖK. Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange. Expert Syst Appl. 2011;38(5):5311–5319.
7. Guresen E, Kayakutlu G, Daim TU. Using artificial neural network models in stock market index prediction. Expert Syst Appl. 2011;38(8):10389–10397.
8. Zahedi J, Rounaghi MM. Application of artificial neural network models and principal component analysis method in predicting stock prices on Tehran Stock Exchange. Phys A. 2015;438:178–187.
9. Qiu M, Song Y, Akagi F. Application of artificial neural network for the prediction of stock market returns: The case of the Japanese stock market. Chaos Solitons Fractals. 2016;85:1–7.
10. Ni LP, Ni ZW, Gao YZ. Stock trend prediction based on fractal feature selection and support vector machine. Expert Syst Appl. 2011;38(5):5569–5576.
11. Kumar D, Meghwani SS, Thakur M. Proximal support vector machine based hybrid prediction models for trend forecasting in financial markets. J Comput Sci. 2016;17:1–13.
12. D. D. Wu, L. Zheng, and D. L. Olson, "A decision support approach for online stock forum sentiment

January - February 2020
ISSN: 0193 - 4120 Page No. 64 - 72

analysis," IEEE Trans. Syst., Man, Cybern., Syst., vol. 44, no. 8, pp. 1077–1087, Aug. 2014.

13. Kirilenko, A, Kyle AS, Samadi M, Tuzun T (2017) The flash crash: high-frequency trading in an electronic market. J Finance.
14. Neil, J, Guannan Z, Eric H, Jing M, Amith R, Spencer C, Brian T (2012) Financial Black Swans driven by ultrafast machine ecology.
15. Guresen, E, Kayakutlu G, Daim TU (2011) Using artificial neural network models in stock market index prediction. Expert SystAppl 38:10389–10397.
16. Schumaker, RP, Zhang Y, Huang C-N, Chen H (2012) Evaluating sentiment in financial news articles. Decis Support Syst 53:458–464.
17. Wong, W-K, Manzur M, Chew B-K (2010) How rewarding is technical analysis? evidence from Singapore stock market. Appl Financial Econ 543–551.
18. Namaki, A, Shirazi AH, Raei R, Jafari GR (2011) Network analysis of a financial market based on genuine correlation and threshold method. Physica A: Stat MechAppl 390:3835–3841.
19. Creamer, GG, Ren Y, Nickerson J (2013) Impact of dynamic corporate news networks on assets return and volatility In: SocComput (SocialCom) 2013 ASE/IEEE International Conference.
20. Sun, X-Q, Shen H-W, Cheng X-Q (2014) Trading network predicts stock price. Sci Reports 4(3711).