Sentiment Analysis of Event Driven Stock Market Price Prediction

Vikrant Kumar Kaushik [1], Arjun Kumar Gupta [2], Ashish Kumar [3], Abhishek Prasad [4], B. Lalitha [5]
[1 – 4] B.Tech Scholar, Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, Tamil Nadu, India.
[5] Assistant Professor (O.G.), Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, Tamil Nadu, India.

Abstract – One of the major investment activities in stock market is trading of stocks. In a volatile market environment, a precise prediction of market trend is very essential. investors developed a number of stock analysis method that could help them predict the direction of stock price movement. Prediction of equity price, based on the current financial statistics, is vital for investors. It helps the investors to know which stock will rise or fall over certain period of time. For this purpose, numerous methods have been devised which have been successful partially. Hence, this proves that process of stock prediction is not an easy task. So in this paper an additional component of measurement is discussed which could be added to the existing methods. The traditional market movement based prediction when combined with the sentiments or opinions of investors and experts will give better results. This ideology has been explored in this paper.

1. INTRODUCTION

Stock market exchange is a finance related institution which allows trading and transaction of various types of commodities (monetary values, precious metals, etc.) among stock broker components. People see stocks as a means to earn a profit considering its turnover of trading reaching billions of US dollars. Trading of goods are done considering theirs value to determine whether profit has occurred or not. In general, the value of a stock is given by its entry on the stock exchange and the volume of its transactions. Greater the number of transactions for a share, the more valuable it is and conversely lesser the volume of transactions, its importance among some traders reduces and its value decreases. This anticipation of the market can generate profits or losses, depending on the power to predict future values. Thus, the problem arises: for a given stock market history, determine the correct moment of buying/entry or selling/exit the goods for maximizing profit.

This aspect has attracted researchers in predicting the values of the goods. Numerous of algorithms in Artificial Intelligence, Machine Learning, tried to solve the problem described above. Some of them are: AR models(Autoregressive), ARIMA (autoregressive integrated moving average), CBR (Case-Based Reasoning), ANN (Artificial Neural Networks), GA (Genetic Algorithm), SVM (Support Vector Machines), SVR(Support Vector Regression) [3],PCA (Principal component analysis). Because of the non-linear nature of the stock market signals, most methods have yet to give promising answers while others have not been very satisfactory on the stock market exchange.l.

2. LITERATURE REVIEW

The companies which are providing the financial services is developing their products to provide the future market price prediction. The stock market needs larges amount of information about the market and one of its area is collection of information about daily price prediction and also stock mining (in case of cryptocurrencies). If some rules could be created to help the investors making good investment decisions in different stock markets according to their performance then the price prediction of stock market becomes increasingly important. One of the algorithm had been adopted by Shin – et –al in 2005 called genetic algorithm; Some rules of trading was generated for the Korean Stock Price Index 200 KOSPI- 200, which is based on a modified kNN to determine where correlated areas fall in the input space which improves the performance of the stock price prediction in the market for the period 1987-96. The models mentioned above were provided in the Zimbabwe stock exchange to predict the stock prices which included Weightless Neural Network (WNN) model and single exponential smoothing (SES) model Mpofu (2004). The methodology of the stocks approach was provided by Gavrilov et al. (2004) to group 750 stocks from the Standard & Poor. The data showed a list of 252 numbers including the opening stock price. A different genetic algorithm was presented by Cao (1977) to discover pair relationship in stock data based on user preferences. This study developed potential guidelines to mine pairs of stocks, stock-trading rules, and markets; and also represented that such approach is useful for real time trading. However, other studies adopted kNN as prediction algorithms such as (Subha et al., 2012; Liao et al. 2010; Tsai and Hsiao 2010; Qian and Rasheed, 2007).

3. RELATED WORK

Algorithm methodologies and concepts:- The kNN algorithm method is basically meant to be used on the stock data. Also, mathematical calculations regarding the concepts and visualization models are provided and discussed below:-

- k-Nearest Neighbor Classifier (kNN)
- Mathematical Calculations and Visualizations Models

a. k-Nearest Neighbor Classifier (kNN)

K-nearest neighbor technique is a machine learning algorithm which is considered as very simple to implement (Aha et al. 1991). The stock price prediction problem can be understood by mapping into a similarity based classification. Both the historical stock data and the data to be tested is mapped into a set of vectors. Each vector represents N dimension for each stock features. Thus, a similarity metric such as Euclidean distance is computed to take the decision.

kNN is always considered as a lazy learning algorithm that does not build a model or function previously, but yields the closest k records of the training data set that have the highest similarity to the data to be tested (i.e. query record). Then, a majority vote is performed among the selected k records to determine the class label and then that class label is assigned to the query record data. The prediction of stock market closing price is computed using kNN as follows:

a) Determination of the number of nearest neighbors, k. b) Computing the distance between the training samples data and the query record data. c) Sorting all the training record data according to the distance values computed above. d) Then using a majority vote for the class labels of k nearest neighbors, and assign it as the prediction value of the query record.

b. Mathematical Calculations and Visualizations Models

This represents an overview of equations that were applied for predicting next day price. The calculations includes error estimation, total sum of squared error, average error, cumulative closing price when sorted using predicted values, k-values and training Root Mean Square (RMS) errors.

a) Root Mean Square Deviation (RMSD) is accuracy metric that computes the differences between the estimated values, Y, and the actual values, X. The total of RMSD is aggregated into a single value measure. RMSD = SQRT(Y-X) 2 .

b) Explained Sum of Squares (ESS) is computed as follows: ESS = Where yi: is the predicted variable, and y being the actual value.

c) Average Estimated Error (AEE) AEE is the total sum of RMS errors for all variables in stock records divided by the total number of the records. Visualization Graph is provided to evaluate the performance of Knn lazy learning model, lift graph is applied and drawn for different companies stock values. The lift chart symbolizes the enhancement that a data mining model offers when distinguished against a random estimation, and the change is expressed in terms of a lift score. Going through the lift scores for a variety of parts of the data set and for different models, it can then be decided that which model is supreme and what percentage of the cases within the data set would gain from employing the predictions model. Furthermore, using the lift chart assists in distinguishing how accurate predictions are for various models with identical predictable characteristic features.It also shows the ratio between the results obtained using the predictive model and not using the predictive model. The other graph is applied where the plot curves to show the relation between the actual and predicted stock price value.

4. PROPOSED MODELLING

The implementation of the proposed system is carried out in three different steps. In the first step, the stock market datasets are obtained and using Markov chain method, the stock market prediction is done. In the second step, user opinions obtained from stock investing websites and banking sites are then used for sentiment analysis using naïve Baye's theorm. The Markov chain prediction and Sentiment analysis steps are carried out independently on different datasets for a given date range. In the final step, the results obtained from the Markov chain method and Naive Baye's are then combined to get the final prediction results.

A) Hidden Markov Chains

Markov chains, named after Andrey Markov, are mathematical systems that hop from one "state" (a situation or set of values) to another. A Markov chain tells you the probability of hopping, or "transitioning," from one state to any other state. To experiment whether the stock market is influence by previous market events, then a Markov model is a perfect experimental tool.

Fig 1: Use case diagram

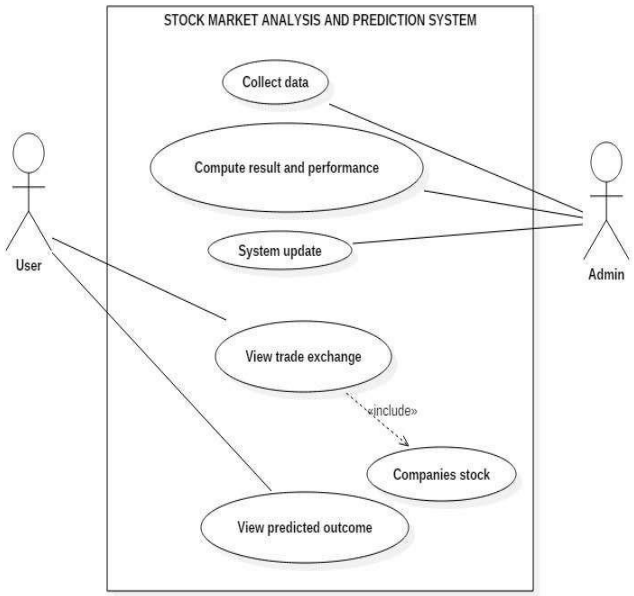
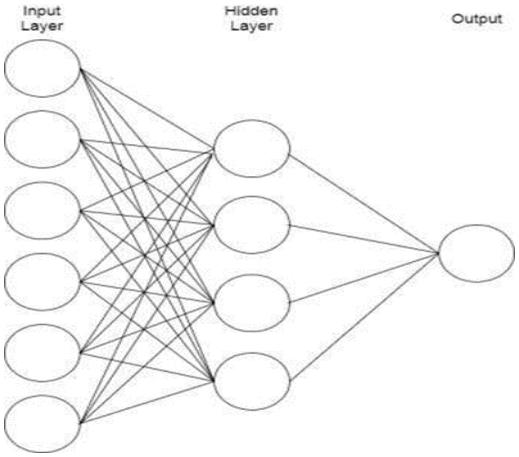


Fig 2: Markov chain model



In the proposed system, the Markov Chain method will be used to collect enough sequences, even of varying lengths, to find patterns in past behaviour of market trends. The following steps are followed to make stock market prediction.

1. In order, get a better understanding of the market's behaviour, This obviously isn't using any of Markov's ideas and is just predicting future behaviour on the basis of an up-down-up market pattern.
2. An approach is to simplify each event within a sequence into a single feature. The value is split into 3 groups - Low, Medium, High. The simplification of the event into three bins will facilitate the subsequent matching between other sequence events and will capture the story so it can be used to predict future behaviour.
3. To better generalize stock market data, collect the percent difference between one day's price and the previous day's. Once collected all of them, they can grouped into groups of equal frequency. all the features for a particular are combined event into a single feature.
4. Now, creating Markov Chains for features like one for 'days' with 'volume jumps' and another for 'days' with 'volume drops'.
5. An important approach of the proposed system is to separate sequences of events into separate data sets based on the outcome. As the system is proposed to predict volume changes, one data set will contain sequences of volume increases and another, decreases. This enables each data set to offer a probability of a directional volume move and the largest probability, wins.
6. A transition matrix is the probability matrix from the Markov Chain. In its simplest form, it is read by choosing the current event on the y axis and look for the probability of the next event off the x-axis.
7. Thus, with the help of transition matrix of Markov chain, we can find the probability for the occurrence of an event and thus market trend prediction can be made in this way.

B) Sentiment Analysis

1. Grammatical tagging

The user opinion datasets obtained from banking and stock investing sites are first passed through the process of grammatical tagging. Grammatical tagging is the process of identifying and mapping words to their parts of speech in English language. The words are categorized based upon their definition and also on relative context to neighboring words. There are eight standard parts of speech in English are adverb, pronoun, noun, interjection, adjective, preposition, articles and conjunction. A given word can belong to more than one category and hence some more categories are created such as Verb Finite, Possessive Adjective etc. The main steps involved in grammatical tagging are

i. Token formation: A given text is taken and broken up into tokens which would be further analysed.

ii. Unknown words look up: An unknown word is categorized using a Lexicon which contains a list of likely parts of speech. Compiler or interpreter works as a lexical analyzer.

iii. Context Resolution: In this process, words which can be placed in multiple categories are resolved based on probability and context of neighboring words.

2. Naive Bayes Classification.

After the words have been classified, a probabilistic classification theorem called is used to classify the text for sentiment classification.

Naïve Bayes classifier is a probabilistic classifier that can be used for sentiment classification in relation to stock data opinions. It uses word frequencies as the characteristic feature to classify text corpuses.

A general formula for Naïve Bayes is given as: -

$$p(f_k/\mathrm{g}) = \frac{[\mathrm{p}(\mathrm{g}/f_k).\ \mathrm{p}(f_k)]}{p[g]}$$

Here *'f_k'*stands for a set of features and *'g'* represents class of features. For a given text review document, let 't' represent the

document, a fixed set of output classes $O = o_1, o_2,..., o_M$ and return a predicted class $o \in O$. Naive Bayes is a probabilistic classifier, meaning that for a document t, out of all classes $o \in O$ the classifier returns the class o which has the maximum probability in given the document. Mathematically, class *'o'* for a given document *'t'* can be represented as: -

$o = argument\ maximum\ of\ o \in O\ \ p(o/t)$

where,

$$p(o/t) = \frac{[p(t/o).\ p(o)]}{p[t]}$$

Since P(d|c)P(c) /P(d) will be computed for each possible class. But *P(d)* doesn't change for each class; computation is done for the most likely class for the same document d, which must have the same probability P(d). Thus, a class can be chosen that maximizes this simple formula: -

$O = argument\ maximum\ of\ o \in O\ P(o/t) = argument\ maximum\ of\ o \in O\ P(t/o)\ P(o)$

With a multinomial event model, samples represent the frequencies with which certain events have been generated by a multinomial (p_1, p_2…..p_n). where p_i is the probability that event i occurs (or K such multinomials in the multiclass case). This is the event model typically used for document classification, with events representing the occurrence of a word in a single document following two main assumptions - bag of words assumption and independence assumption.

The first is the bag of words assumption: It is assumed that position doesn't matter, and that any occurrence has the same effect on classification whether it occurs as the 1st, 20th, or last word in the document. Thus it can be assumed that the features $f_1, f_2..., f_n$ only encode word identity and not position.

The second is commonly called the Naive Bayes assumption: It is an independence assumption that the probabilities $P(f_i/o)$ are independent given the class o and hence can be 'naively' multiplied as follows:

$P(f_1, f_2,...., f_n/o) = P(f_1/o)\cdot P(f_2/o)\cdot ... \cdot P(f_n/o).$

The final equation for the class chosen by a naive Bayes classifier is thus:

$NBC = argument\ maximum\ of\ o \in O\ P(o)\ \prod f \epsilon F\ \ P(f/o).$

5. MODULE DESCRIPTION

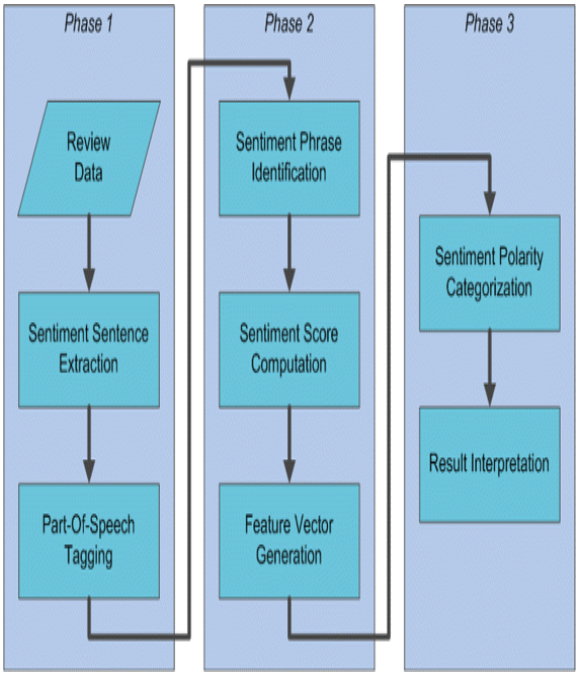
Hidden Markov Chain Modules –

Sequence generation – In this module, the data is broken up into samples of sequences leading to different patterns. Different patterns generated are matched to predict future market trends.

Grouping of sequence events with equal frequencies – As an approach to generalize stock market information data, this module finds the similarity between patterns and groups them into categories having equal frequencies.

Creation of Markov Chains and probability calculation – In this module, Markov chains from equal frequency groups and then a transition matrix or probability matrix used to make predictions. It reads the values of current event on one axis and probability is found on the other axis.

Sentiment Analysis Modules –



Sentiment Sentence Extraction - Extraction of key sentences with sentiments from text discourse forms an important role in analysis of sentiments. The technique used is Multiple Source Features (MSF). In MSF technique, for each sentence, four sources of features are designed which are, lexical sentiment, global position, word grammar indicator, and title similarity. After this, these features are combined linearly to get a score that indicates the probability that the sentence is a key sentiment sentence.

Part of Speech Tagging - Part-of-speech tagging or word-category disambiguation, refers to the process of assigning a word in a corpus as corresponding to a particular part of speech which is based on both its definition and its context i.e., its relationship with adjacent and related words in a phrase, sentence, or paragraph. In the proposed system Two algorithms will be used two algorithms for this, Brill Tagger and Baum Welch. Brill tagger algorithm can be summarized an "error-driven transformation-based tagger". It is a type of supervised learning, which aims to minimize error through a

transformation based process where it is assumed that a tag is assigned to each word which is changed using a set of predefined rules.

Sentiment phrase Identification – Sentiment phrase Identification module simply combines the results obtained from sentence extraction module and Part-Of-Speech module to get phrase from the sentence which would comprise its grammatical information as well.

Sentiment Score Calculation – Sentiment Score Calculation is based on calculating a sentiment score for each sentence so as to find out how positive or negative is any posted message. Calculation of scores can be done in several ways. Here a simple yet useful approach is applied to define a score formula.

Score = Number of positive words - Number of negative words

A Score > 0 concludes that the sentence has an overall 'Optimistic' expression.

A Score < 0 concludes that the sentence has an overall 'Pessimistic' expression.

A Score = 0 concludes that the sentence has an overall 'Neutral' expression.

Feature Vector Generation - Feature Vector Generation is done using machine leaning conepts. Symbolic techniques or Knowledge base approach along with Machine learning techniques are the two main techniques used in sentiment analysis. Knowledge base approach requires a large database of predefined emotions and an efficient knowledge representation for identifying sentiments. Machine learning technique uses a training set to develop a sentiment classifier that classifies sentiments. Since a predefined database of entire emotions is not required for machine learning approach, it is rather simpler than Knowledge base approach.

Sentiment Polarity Categorization – This module includes ways to classify the sentences broadly in to three category – positive, negative and neutral. By using the property that sentiment classification has two opposite class labels (i.e., positive and negative), first propose a data expansion technique by creating sentiment reversed reviews. The original and reversed reviews are constructed in a one-to-one correspondence. On analysis of current dual sentiment analysis we propose our approach to analysis sentiment as well as its automatic rating count. This can be calculated by using user review on the basis of positive, negative and neutral response. Than calculate all review and display the result analysis.

6. RESULTS AND DISCUSSIONS

The proposed system of stock market prediction using a technique which combines the Markov chain based prediction and sentiment analysis of user opinion has led to improvement in stock market analysis and prediction up 3 percent.

7. CONCLUSION

The aim of study is to evaluate the performance for sentiment classification in terms of accuracy, precision and recall. In this paper, we combined two independent algorithms - supervised machine learning algorithms of Naïve Bayes' and mathematical Hidden Markov Chain method. The experimental results show that the stock market prediction made by Hidden Markov Chain combined with the market trend prediction by made by Naïve Bayes sentiment analysis yields a better prediction of stocks then the existing techniques and thus can be implemented by stock markets and banks to benefit investors, companies and to improve stock market performance

REFERENCES

[1] K. Schierholt and C. Dagli, Stock Market Prediction Using Different Neural Network Classification Architecture, Computational Intelligence for Financial Engineering 1996
[2] Aditya Gupta and Bhuwan Dhingra, Stock Market Prediction Using Hidden Markov Models, IEEE 2012.
[3] Thien Hai Nguyen, Kiyoaki Shirai, Julien Velcin [2015], "Sentiment analysis on social media for stock movement prediction."
[4] Bollen, J., Mao, H., & Zeng, X. [2011], "Twitter mood predicts the stock market."
[5] Gamon M. Sentiment classification on customer feedback data: Noisy data, large feature vectors, and the role of linguistic analysis. // ACM, 2004.
[6] G. Vinodhini and R. Chandrasekaran, "Sentiment analysis and opinion mining: A survey," International Journal, vol. 2, no. 6, 2012.
[7] Rishabh Soni, K. James Mathai [2015], "Improved Twitter Sentiment Prediction through 'Cluster-then-Predict Model'."
[8] E. Boiy, P. Hens, K. Deschacht, and M.-F. Moens, "Automatic sentiment analysis in on-line text," in Proceedings of the 11th International Conference on Electronic Publishing, pp. 349–360, 2007.
[9] Z. Niu, Z. Yin, and X. Kong, "Sentiment classification for microblog by machine learning," in Computational and Information Sciences (ICCIS), 2012 Fourth International Conference on, pp. 286–289, IEEE, 2012.
[10] X. Wu, M. Fund and A. Flitman, "Forecasting Stock Performance using Intelligent Hybrid Systems", Springerlink, 2001, pp. 4 4 7-4 56.
[11] Song yanxue, Zhang shaowu, and Lin hongfei. Sentence Sentiment analysis based on ambiguous words. 2012. May, Vol. 26, No. 3. Pp. 38-42.
[12] S. Li and C. Huang, "Sentiment classification considering negation and contrast transition," Proceedings of the Pacific Asia Conference on Language, Information and Computation (PACLIC), 2009.
[13] S. Li, R. Xia, C. Zong and C.Huang, "A framework of feature selection methods for text categorization," Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), pp. 692-700, 2009.
[14] F.Luo, J. Wu and K. Yan, A Novel Nonlinear Combination Model Based on Support Vector Machine for Stock Market Prediction, Proc. Of the 8 th World Congress on Intelligent Control and Automation, July 6-9 2010, Jinan, China.
[15] M.R. Hassan. "A combination of hidden markov model and fuzzy model for stock market forecasting". Journal of Neurocomputing, pages 3439–3446, 2009.
[16] L.R. Rabiner. "A tutorial on hidden markov models and selected applications in speech recognition". Proceedings of the IEEE, pages 257– 286, 1989.
[17] "Hidden markov models and the baum-welch algorithm". IEEE Information theory society newsletter, Dec 2003.

[18] Md. Rafiul Hassan, Baikunth Nath and Michael Kirley, " A fusion model of HMM, ANN and GA for stock market forecasting," Expert systems with Applications., pp. 171-180,2007.
[19] Li, X.; Parizeau, M.; Plamondon, R. Training Hidden Markov Models with Multiple Observations—A Combinatorial Method. IEEE Trans. PAMI 2000, 22, 371–377

Authors

Vikrant Kumar Kaushik - Under graduate Student, Computer Science and Engineering at SRM Institute of Science and Technology Ramapuram Chennai(TN)

Arjun Kumar Gupta - Under graduate Student, Computer Science and Engineering at SRM Institute of Science and Technology, Ramapuram Chennai(TN)

Ashish Kumar - Under graduate Student, Computer Science and Engineering at SRM Institute of Science and Technology, Ramapuram Chennai(TN)

Abhishek Prasad - Under graduate Student, Computer Science and Engineering at SRM Institute of Science and Technology Ramapuram Chennai(TN)

B.Lalitha and Assistant Professor (O.G), Computer Science and Engineering at SRM Institute of Science and Technology Ramapuram Chennai(TN)