International Journal of Global Engineering (IJGE) E- ISSN: 2456-3099

VOL 2 ISSUE 4 (2018) PAGES 28 - 34

Received: 25/03/2018. Published: 06/04/2018

Stock Market Prediction Using Machine Learning Techniques

[1]Poonkodi.M, [2]Swathi.P, [3]Monisha.E and [4]Rohit.A

[1]Assistant Professor, [2,3,4]UG Scholars

Department of Computer Science, SRM Institute of Science and Technology

Vadapalani Campus, Chennai, India

Corresponding Author: achantarohit@gmail.com

Abstract— Predicting stock market movements is a well-known problem of interest that every stock trader would look to gain advantage in the stock market business. Now-a-days social media is perfectly representing the public sentiment and opinion about current events. Especially, Twitter has attracted a lot of attention from researchers for studying the public sentiments. This project focuses on efficient Stock Market Prediction using Machine Learning Techniques. In this system, Sentiment Analysis and Supervised Machine Learning Principles have been applied to the tweets extracted from Twitter and analyze the correlation between stock market movements in a company and sentiments in tweets. The existing system includes analysis based upon stock investors sentiments obtained through message board comments where polarity calculation is used to track sentiment. The proposed system aims to study Candle Stick Pattern along with the graphical convention of the Support Vector Machine for better accuracy.

Index Terms— Stock market, Public Sentiment, Prediction, Machine Learning, Analysis, Polarity Calculation, Candle Stick Pattern, Support Vector Machine

I. INTRODUCTION

In the finance field, stock market and its trends are extremely volatile in nature. It attracts researchers to capture the volatility and predict its next moves. Investors and market analysts study the market behavior and plan their buy or sell strategies accordingly. The program focuses on estimating the stock price of the company using the investor's comments extracted from Twitter. It is expected that comments extracted from social media have a strong impact on a company's stock price. In this paper, we have investigated the possibilities of analyzing social media with Machine Learning and Sentiment Analysis for stock market forecasting. We calculate the polarity for each of the words in the posts made on Twitter regarding a company based on the sentiment. A positive post gets a positive value and a negative post gets a negative value. The overall polarity is calculated for each post by adding up the individual polarities. These polarities are plotted in a histogram or a pie chart to visually represent the data Collected. The representation helps us interpret and predict the trend or the movement of the stocks for a particular company making use of real time data in the form of Twitter posts.

Mainly there are two methods for forecasting market trends. One is Technical Analysis and the other is Fundamental Analysis. Technical Analysis considers past price and volume to predict the future trend whereas Fundamental Analysis on the other hand, involves analyzing its financial data to get some insights. The efficacy of both Technical and Fundamental Analysis is disputed by the Efficient-Market Hypothesis which states that stock market prices are essentially unpredictable. This research follows the Fundamental Analysis Technique to discover the future trend of a stock by considering news articles about a company as prime information and tries to classify news as good (positive) and bad (negative). If the news sentiment is positive, there are more chances that the stock price will go up and if the news sentiment is negative, then stock price may go down. This research is an attempt to build a model that predicts polarity which may

International Journal of Global Engineering (IJGE) E- ISSN: 2456-3099
VOL 2 ISSUE 4 (2018) PAGES 28 - 34
Received: 25/03/2018. Published: 06/04/2018

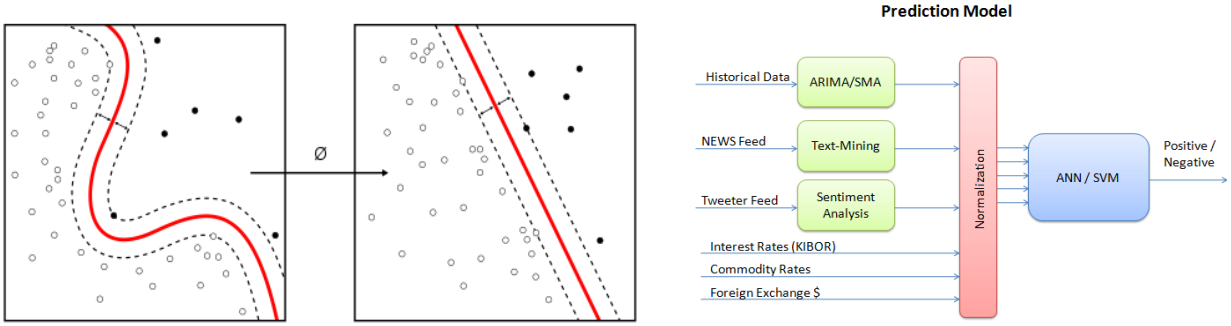
affect changes in stock trends. In other words, check the impact of tweets on stock prices. Supervised Machine Learning as classification and other Text Mining Techniques have been used to check the polarity and also be able to classify unknown tweets, which is not used to build a classifier. Three different classification algorithms are implemented to check and improve classification accuracy. Investor's tweets are extracted from Twitter for predicting the stock price of the company. The result implies that there exists a strong correlation between rise and fall of stock price with the investor's sentiment in tweet.

II. METHODOLOGY

The initial step is to gather information regarding companies through Social Media. It requires extracting certain feature of the said Social Media. Feature Extraction starts from an initial set of measured data and builds derived values (features) intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps, and in some cases leading to better human interpretations. When the input data to an algorithm is too large to be processed and it is suspected to be redundant (e.g. the same measurement in both feet and meters, or the repetitiveness of images presented as pixels), then it can be transformed into a reduced set of features (also named a Feature Vector). Determining a subset of the initial features is called Feature Selection. The selected features are expected to contain the relevant information from the input data, so that the desired task can be performed by using this reduced representation instead of the complete initial data. Feature Extraction is a general term for methods of constructing combinations of the variables to get around these problems while still describing the data with sufficient accuracy. It involves reducing the amount of resources required to describe a large set of data. When performing analysis of complex data one of the major problems stems from the number of variables involved. We extract the posts regarding the Companies to determine trend in their stocks. The data obtained is processed using the following algorithms

A. Support Vector Machine

Support Vector Machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). In this case, we calculate whether a given post is positive or negative. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and are predicted to belong to a category based on which side of the gap they fall into.



International Journal of Global Engineering (IJGE) E- ISSN: 2456-3099
VOL 2 ISSUE 4 (2018) PAGES 28 - 34
Received: 25/03/2018. Published: 06/04/2018

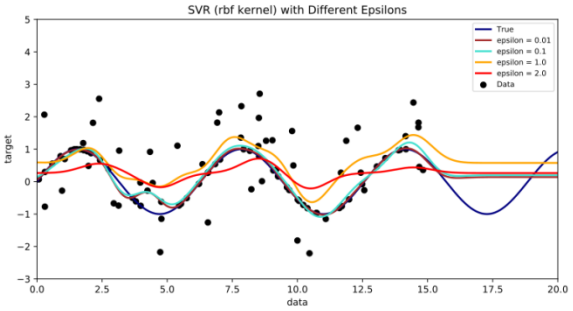
In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the Kernel Trick, implicitly mapping their inputs into high-dimensional feature spaces.

When data are not labeled, supervised learning is not possible, and an unsupervised learning approach is required, which attempts to find natural clustering of the data to groups, and then map new data to these formed groups. However, we label the data of our classifier as Positive and Negative where the word of a post are mapped accordingly. The Support Vector Clustering algorithm created by Hava Siegelmann and Vladimir Vapnik, applies the statistics of support vectors, developed in the Support Vector Machines algorithm, to categorize unlabeled data, and is one of the most widely used clustering algorithms in industrial applications. Implenting a clustering algorithm may prove useful to predict future trends of a given company.

B. Support Vector Regression

Support Vector Regression (prediction) with different thresholds ε. As ε increases, the prediction becomes less sensitive to errors. It might be comparatively effective to not only classify words but also to predict the next likely point on the plane.

A version of SVM for regression was proposed in 1996 by Vladimir N. Vapnik, Harris Drucker, Christopher J. C. Burges, Linda Kaufman and Alexander J. Smola. This method is called Support Vector Regression (SVR). The model produced by support vector classification (as described above) depends only on a subset of the training data, because the cost function for building the model does not care about training points that lie beyond the margin. Training data here is the posts extracted from Social Media. Analogously, the model produced by SVR depends only on a subset of the training data, because the cost function for building the model ignores any training data close to the model prediction. Another SVM version known as least squares support vector machine (LS-SVM) has been proposed by Suykens and Vandewalle.



III. PROPOSED METHOD

The technique proposed in this research is to use different factors impacting the market as input attributes for the model. The output of the model is one of the two defined classes that are Positive Market and Negative Market. Fig. shows the structure of the proposed model. The model studies and compares the results of all four machine learning algorithms defined above. All the attributes that are used in this model were continuous numeric values and were of different range. These attributes are therefore normalized between [+1,-1] because all the parameters used can have positive and negative values. Each attribute will be discussed separately.

International Journal of Global Engineering (IJGE) E- ISSN: 2456-3099
VOL 2 ISSUE 4 (2018) PAGES 28 - 34
Received: 25/03/2018. Published: 06/04/2018

A. Factors

Following are different factors that were found to have some impact on market performance in different studies.

1. Market History

The first attribute used as an input for the model is the historical closing index of KSE-100. The historical data was not made part of the model directly but after applying statistical techniques including ARIMA and SMA over the data. The window size used is 4.

2. The NEWS

NEWS is another influencing factor considered for market performance. NEWS can be of different categories but in this model only business, financial, political and international event based NEWS were included.

3. General Public Mood

The market performance greatly depends on the investor's mood and sentiment. The collective sentiment of people may drive stock market performance. This can be achieved with the help of social media. In this research, Twitter is used as the source of public sentiment.

4. Commodity Price

The changes in prices of different commodities do have an impact over the market behavior. Change in price of commodities like petrol reflects on almost all items. Commodities including Gold, Silver and Petrol are used as inputs to the model.

5. Interest Rate

The interest rates issued by State Bank of Pakistan to all the banks that provides loan to their customers also have an effect on the market. The Karachi Inter Bank Offer Rate (KIBOR) is issued on daily basis for different durations. In this study, 1-week rates are used.

6. Foreign Exchange

Change in Foreign Exchange rate has been assumed to affect the market performance by many. Historical exchange rate between the Pakistan Rupee (PKR) and the US Dollar (USD) was used as an input to the model. On basis of the above factor, a total of 9 parameters (i.e., Oil rates, Gold Rates, Silver Rates, FEX, SMA, ARIMA, KIBOR, NEWS, Twitter) are used as inputs for the prediction model, whereas the two classes are designed to be output by the mode.

B. Scope

The data used in this study spreads over the time range the user wishes to choose. The data gathered from NEWS and Twitter lie between current day close till the next day's market opening.

C. Text Mining

NEWS and Twitter data were available in the form of feed which was processed using Text Mining Techniques. The library Opinion Finder was used for this purpose. In Twitter feed, the use of non-English words was very frequent. This could cause wrong classification of the text. The work was done to implement the dictionary that translates an Urdu word written in Roman to its alternate English word. Later this updated feed was sent to feed processor application that

VOL 2 ISSUE 4 (2018) PAGES 28 - 34
Received: 25/03/2018. Published: 06/04/2018

reads the whole text and classifies it into one of the two classes.

IV. ARCHITECTURE DESCRIPTION

A. Authentication

The secret keys and authentication number generated by Twitter API are given as input. A new authentication window opens rendering the final token for authentication. Once this is entered, the access for fetching tweets is granted.

B. Data Fetch

Enter the Stock name or ID and specify the number of tweets to be fetched. The fetched tweet array is converted into a list along with index numbers. The raw form of tweets is fetched and displayed in R console.

C. Data Clean

Data cleaning and segmentation is performed to remove unwanted errors in analysis. Emoticons, grammatical mistakes and unwanted URLs are removed in this phase. Processed data is stored in a data structure. Text data is unstructured data. So, we cannot provide raw test data to classifier as an input. Firstly, we need to tokenize the document into words to operate on word level. Text data contains more noisy words which are not contributing towards classification. So, we need to drop those words. In addition, text data may contain numbers, more white spaces, tabs, punctuation characters, stop words etc. We also need to clean data by removing all those words. For this purpose, we created an own stop-word list which specifically contains stop words related to finance world and also general English stop words.

D. Database Connection

Natural Language Processing is performed and positive and negative words are stored in different files. This data list is mapped and connected to the R console. The positive and negative words are retrieved into different separate data structures. Further words are to be added through the R console.

E. Score Calculation

Depending on the frequency of occurrence of positive or negative words the positive and negative score is calculated. The net polarity per sentence is given by, positive score- negative score. The total percentage of positivity and negativity of all retrieved can also be predicted.

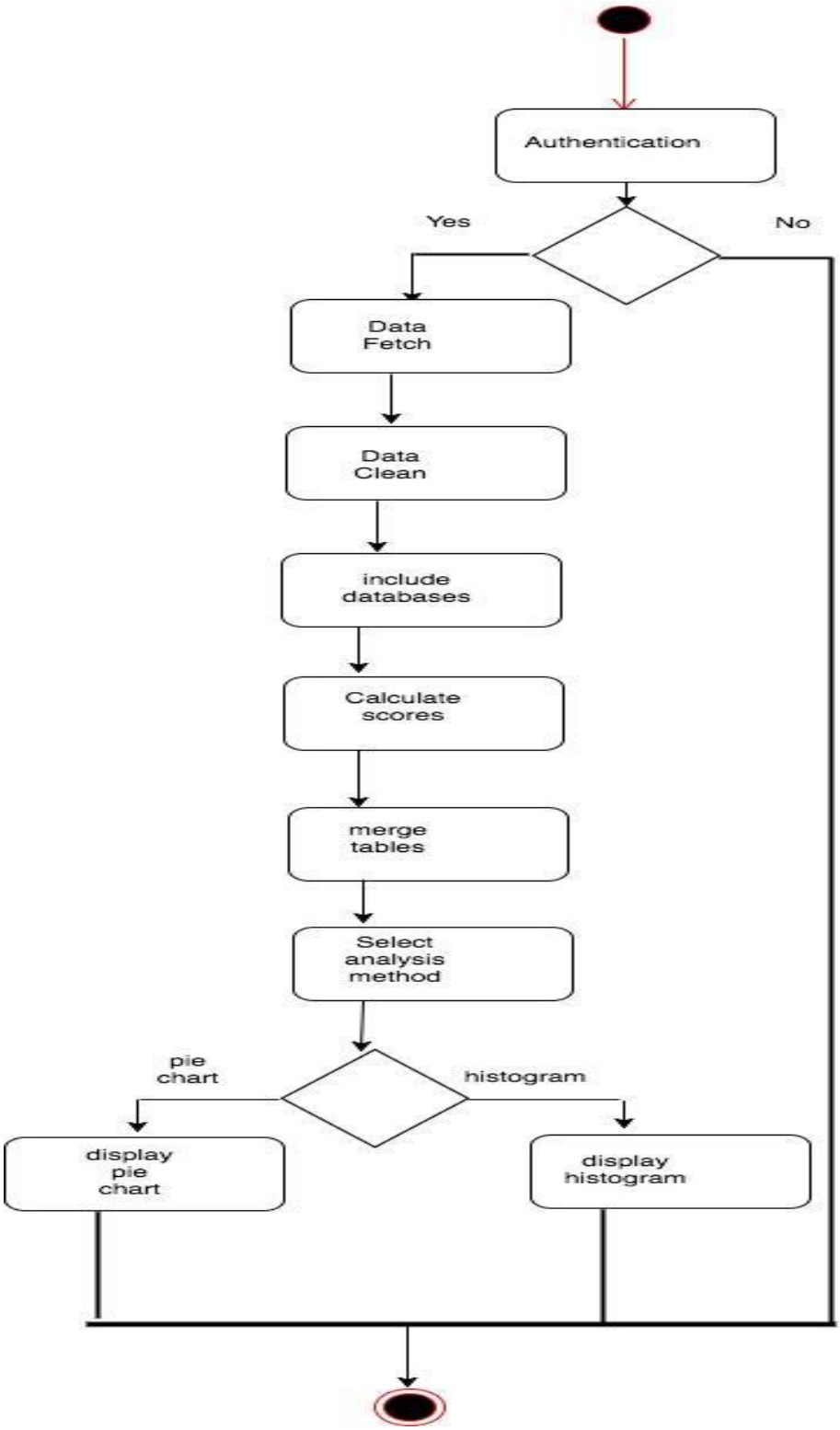
F. Merge Score and Table Construction

The tweet index, net polarity, positive and negative polarity is mapped and a table is constructed by merging the positive score table and the negative score table.

G. Graphical Analysis of Prediction

The scores are plotted in a graphical form for analysis of market prediction. The expected range of rise or fall of the stock is depicted. Histograms and pie charts are developed for this purpose.

International Journal of Global Engineering (IJGE) E- ISSN: 2456-3099
VOL 2 ISSUE 4 (2018) PAGES 28 - 34
Received: 25/03/2018. Published: 06/04/2018



V. CONCLUSION AND FUTURE WORK

H. Conclusion

In general, the problem of stock market prediction is very challenging and very high accuracies are not easily achievable. Nevertheless, machine learning techniques can provide reasonable

market movement predictions that can be used by investors. The calculated results show that using support vector machines with Gaussian kernel and regularization outperforms the logistic regression and SVM with other kernels. In this study, we were able to get a prediction with very good accuracy. We should emphasize that the current dataset is very valuable, and many extensions to the current algorithm can be applied to it.

I. Future Work

First of all, we could expand the feature space to higher dimensions, and try other feature selection methods such as wrapper model feature selection. Secondly, as a natural extension to our work, one can predict the value of the price jump, rather than just the sign of it. Moreover, one can categorize companies based on their mutual influence on each other, and find an individual learning parameter, θ, for each category (e.g., IT companies, energy related companies, health care companies, etc.). Additionally, for each feature, one can compute the correlation matrix of that feature between different companies. This can be used as an extra information for prediction purposes.

VI. References

[1] Yan Guo, Huifeng Tang, Linhai Song, Yu Wang, Guodong Ding, "ECON: An Approach to Extract Content from Web News Page", 12th International Asia-Pacific Web Conference, 2010.
[2] Y. Wu and F. Ren "Learning sentimental influence in twitter " in Future Computer Sciences and Application (ICFCSA) 2011 International Conference on pp. 119-122 IEEE 2011.
[3] A. Pak and P. Paroubek "Twitter as a corpus for sentiment analysis and opinion mining " in Proceedings of LREC vol. 2010.
[4] I. Ahmed A. Aziz "Dynamic approach for data scrubbing process" (IJCSE) International Journal on Computer Science and Engineering 2010 (ISSN0975–3397).
[5] Y. Hu R. Lu X. Li J Duan and Y. Chen. "esearch on language modeling basedsentiment classification of text" Journal of Computer Research and Development vol. 44 no. 9 pp. 1469-1475 2007.