

Erudition and Contemplation of service for stock market using Data Mining

SRINADH UNNAVA¹

¹ Assistant Professor,
Department of Information Technology,
Sasi Institute of Technology and Engineering (A), Tadepalligudem,
Jawaharlal Nehru Technological University Kakinada, Andhra Pradesh, India

Yaswanth Sai², Ramji Adapa³, Diyya Rayavarapu⁴, Akhil Thota⁵

^{2,3,4,5} IV B.Tech Students,
Department of Information Technology,
Sasi Institute of Technology and Engineering (A), Tadepalligudem,
Jawaharlal Nehru Technological University Kakinada,

Abstract - We surveyed existing and new tactics of stock market prediction. We have taken different approaches i.e Technical Analysis, and the application of Machine Learning. We find evidence in support of the weak form of the Efficient Market Hypothesis, that the previous predicted price does not contain any beneficial facts but out of sample data may be predictive. We show that Data mining and Machine Learning could be used to lead the shareholder decisions. In addition to this we are also using Sentiment Analysis on the basis of product demandable season. Based on our findings, algorithmic trading programs are developed and delivered for end user using Android Application.

Keywords: Sentiment Analysis, Data Mining, Machine Learning, Stock market

1. Introduction

The prediction of stock costs has perpetually been a tough brain-teaser. It's been discovered that the stock value of any company has no direct relation on the economic scenario of the country. Currently stock costs square measure affected thanks to several reasons like company connected news, political events, natural disasters, etc.

The quick processing of those events with the assistance of improved technology and communication systems has caused the stock costs to fluctuate in no time. To speculate cash within the exchange we want to own a thought whether or not the costs of stocks square measure attending to increase or decrease on subsequent day.

There are two approaches of predicting stock exchange behavior. The primary one is predicated on the prediction of future worth values of a stock. Sometimes this needs treating the historical knowledge as statistic associated feeding it to the distinct time-frame signals to an algorithmic rule and attempting to model the long run time points within the signal. The second one is predicated on predicting the long run worth direction of a stock, i.e. estimation whether the worth can rise or fall by the following day, or in a very number of days (trend forecasting).

2. Literature survey

Mantyl'a, Graziotin, and Kuutila proposed a linguistic approach to the problem of sentiment analysis. The object of sentiment analysis has usually been a product or a service whose review has been created public on the web. The first relates to techniques to automatically generate sentiment lexicons. The second relates to systems that analyze sentiment for entire documents. The authors were able to acquire up to 81% of accuracy on their testing dataset. However, the tactic showed a nasty performance with 3 categories("Positive", "Negative", "Neutral") [1].

Sentiment analysis depends on our ability to spot the sentimental terms in an exceedingly corpus and their orientation.[2] We depend separate lexicons for every of seven sentiment dimensions (general, health, crime, sports, business, politics, media). We hand-picked these dimensions supported our identification of distinct news spheres with distinct standards of opinion and sentiment.

Particle Swarm optimization formula (PSO) could be a comparatively recent heuristic search methodology that springs from the behavior of social teams like bird flocks or fish swarm. PSO moves from a group of points to a different set of points in a very single iteration with possible improvement employing a combination of settled and probabilistic rules.[3]

Based on the vendor's feedback we have to calculate positive words, Negative words. The words may be varied but the meaning will not varies (i.e the same word may be adjective, verb) [4]. The project must be delivered to the end-user in a clean and transparent manner. So we choose to deliver in the form of an Android Application.[3] For that we have studied and made work on Android SDK tools and AVD Manager.

We studied how the "Sentiment Analysis" is working on the Twitter comments using Opinion Mining. Using Twitter API we tend to collected a corpus of text posts and shaped a dataset of 3

classes: positive sentiments, negative sentiments, and a collection of objective texts (no sentiments).[5]

The study aims to explore the strength of causative relationship between stock worth search interest and real stock exchange outcomes on worldwide equity market indices.

3. Proposed system

Our proposed system is given in fig 1 is divided into two models. The first model contains three algorithms for predicting the stock price. Then the second model includes sentiment analysis which will be done by using client feeds. The complete integration of this model gives better outcomes when compared to current single isolated system. The model is as follows:

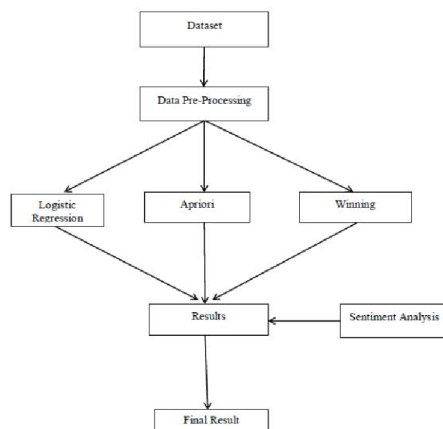


Figure. 1: Proposed System

To implement the first model of the proposed system. These are the following steps:

- Collect the data and develop the dataset for the prediction. This system can work on the different stocks. We have chosen 5 stocks under IT (Information Technology) sector for prediction: Tata Consultancy Services(TCS), Infosys, Tech Mahindra, Mind tree, Wipro.
- With immediate effort preprocessing the data. The Data will be preprocessed using "Logistic regression" to remove the redundancy as well as for filling the missing values. This process is done in "Python" language.
- After preprocessing the dataset implementing the prediction Algorithms. Stock Market may gives some very uncertain results. But that can also be influenced by many other parameters. So we have implemented this on the basis of three prediction Algorithms: KNN, Apriori, Winning was done separately.

The KNN algorithm is a popular method for classifying objects based on closest training examples in the feature space. The following are the parameters in the dataset.

Open value. Closed
Value. High Price.

Low price.
Volume.

To implement prediction we need to split our dataset into "Training" set and "Testing" set which is preferably 70:30. After deciding the number of clusters (K) in which training set is divided. The value of k is taken as sort (no. of data points in training set). Now the training is done using training dataset using the fit method and then extrapolate using the predict method. This comparison gives around 60% of accuracy.



Figure. 2: Results of KNN

As stocks are volatile, to predict more accurate result, another algorithm called Apriori algorithm is employed during this system. Apriori Algorithm uses the open value and the adjusted close value as the parameters for prediction.

Moving on this system also provide opening price of next day using the Winning algorithm which will complement the apriori algorithm output as user can check if the range is achievable or not.

4. Experimental Results

KNN algorithm is applied on the information for an amount of five years to predict the stock value and compare it with the original movements for an amount of five years.

It turns out that the efficiency of the test data is around 65-70% if the data is not largely awry else the accuracy is around 48 – 53%. Experimental results of KNN algorithm on Tata Consultancy services(TCS) stock are as follows:

The accuracy for TCS is around 58%. It seems that there's a necessity for improvement as this prediction is extremely volatile and can't be completely trusty by the users.

In search of better accuracy another algorithm – Apriori algorithm is used. Apriori algorithm provides 5 best buying and shorting values as follows:

This algorithmic rule shows the five values that associate capitalist will place confidence in to perform trades within the order of best to worst as shown in Fig 3. Here the output is given as a variety within which values will fluctuate if happens so; the investor can have 2nd or 3rd values as their best shot.

To more enhance/ support the claims of Apriori, Support Vector Regression (SVR) is employed that provides the open worth for any future. SVR [4] is a regression algorithm. Following is the functions or parameters on which SVR is performed.

- Kernel: The function wants to map lower dimensional information into higher dimensional information.
- Hyper Plane: In SVM [4] basically this is the separation line between the data classes. Although in SVR we are going to define it as the line this will help us to predict the continuous/target value.
- Boundary line: In SVM there are two lines other than Hyper Plane which creates a margin. The support vectors can be on the Boundary lines or outside it. This boundary line separates the two classes. In SVR the concept is same.

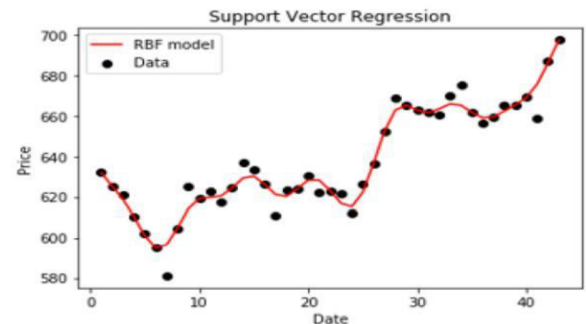
Support vectors-Data points which are considered to be nearest to boundary. They can be nearer or far away from regression line. SVR is implemented on the TCS stock and results are shown in Figures.

The above figure shows the graphical representation between the actual values and the predicted values for Infosys for last 5 years. The Score tells us how well the model fit predict truth values.

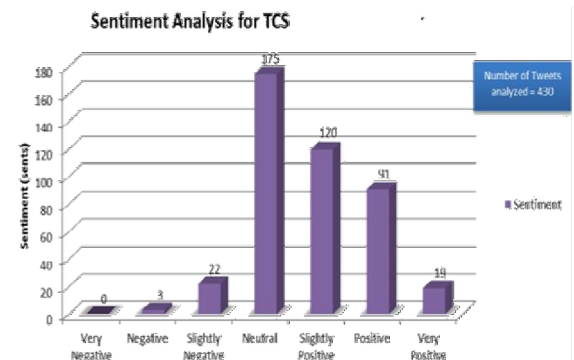
5. Sentimental Analysis

As the stock market results are very much influenced by this market new and lots of different factors, therefore a secondary model of Sentimental Analysis is connected with the system.

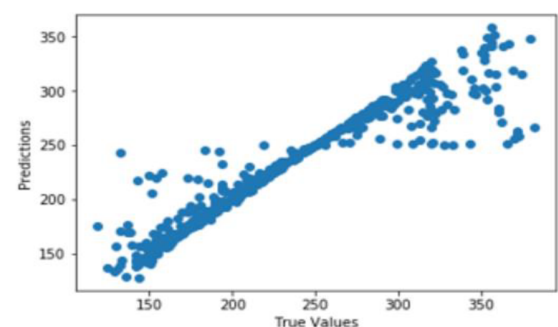
Sentimental analysis can be formed by using various sources of news such as social media, print media etc. Sentimental analysis on twitter feeds is by far the most popular method.



The stock open price for 45th day:
('RBF kernel: ', '702.0757204404197')



The Best 5 Chromosomes When Buying
 min: -0.121565 | max: 0.142286 | previous min: -0.064128 | previous max: 0.388934 | score: 256.500000
 min: -0.039158 | max: 0.173327 | previous min: -0.075932 | previous max: 0.255326 | score: 187.600000
 min: -0.198074 | max: -0.011042 | previous min: -0.051346 | previous max: 0.012216 | score: -7.900000
 min: -0.305503 | max: -0.019589 | previous min: -0.085515 | previous max: 0.161589 | score: -18.600000
 min: -0.395914 | max: 0.046803 | previous min: -0.131716 | previous max: 0.203541 | score: -22.600000
 The Best 5 Chromosomes When Shorting
 min: -0.269975 | max: 0.222943 | previous min: -0.223791 | previous max: -0.004557 | score: 411.700000
 min: -0.216938 | max: 0.002055 | previous min: -0.246327 | previous max: 0.115563 | score: 290.500000
 min: -0.157945 | max: 0.050045 | previous min: -0.157170 | previous max: 0.135180 | score: 118.000000
 min: -0.068127 | max: 0.041252 | previous min: -0.163128 | previous max: 0.000577 | score: 94.100000
 min: -0.748551 | max: 0.115185 | previous min: -0.224166 | previous max: 0.123382 | score: 88.500000



```
print "Score:", model.score(X_test, y_test)
```

Score: 0.8896263441768556

Figure. 5: Sentiment Analysis of TCS

These data has been collected for twitter and feeds are continuously monitored for keywords.

6. Conclusion

The algorithms that are enforced complement one another i.e. if the results are analyzed along higher investment plans will be generated. Conjointly the sentimental model on its own provides basis for the investment.

The accuracy of the system overall is around 70-75% that is ostensibly cheap in such a volatile atmosphere.[7]

References

- [1]. Farah Benamara et al. "Sentiment analysis: Adjectives
- [2]. and adverbs are better than adjectives alone". In: (2007), pp. 1–7.
- [3]. Namrata Godbole, Manja Srinivasaiah, and Steven Skiena. "Large-Scale Sentiment Analysis for News and Blogs." In: *Icwsn* 7.21 (2007), pp. 219–222.
- [4]. Aditya Gupta and Bhuwan Dhingra. "Stock market pre-diction using hidden markov models". In: (2012), pp. 1– 4.
- [5]. Osman Hegazy, Omar S Soliman, and Mustafa Abdul Salam. "A machine learning model for stock market prediction". In: *arXiv preprint arXiv:1402.7351* (2014).
- [6]. Suprabha Kr et al. "Predictive modeling of stock indices closing from web search trends". In: *arXiv preprint arXiv:1804.01676* (2018).
- [7]. Mika V Mantylä," Daniel Graziotin, and Miikka Kuu-tila. "The evolution of sentiment analysis—A review of research topics, venues, and top cited papers". In: *Computer Science Review* 27 (2018), pp. 16–32.
- [8]. Bin Peng, Jinming Yue, and Chen Tianzhou. "The An-droid Application Development College Challenge". In: (2012), pp. 1677–1681.