

Sentiment Analysis with Conventional Data Mining for Predicting Share Market Values

Rohan Kapre¹, Rohan Jawale², Akshay Dhole³, Manisha Sonawane⁴

^{1, 2, 3, 4}Computer Engineering, Shivajirao S Jondhale COE, Mumbai University, India

Abstract: *Proposed system aims to predict stock price movement more accurately by emulating instinctual reasoning by implementing sentiment analysis. Stock market has usually been the proving grounds for data mining applications. There is a lack of efficient models which can imitate the heuristic reasoning of humans based on current events/trends. The proposed system is an attempt to reconcile computed sentiments alongside traditional/more common data mining. Datasets consisting of historical data as well as recent headlines will be mined to predict future stock price.*

Keywords: Sentiment Analysis, data mining, Historical Data, Prediction, Stock Market, Stock Value

1. Introduction

Stock market prediction is the act of trying to determine the future value of a company stock or other financial institution. The successful prediction of a stock's future price could yield significant profit. The efficient-market hypothesis suggests that stock prices reflect all currently available information and any price changes that are not based on newly revealed information thus are inherently unpredictable. Others disagree and those with this viewpoint possess myriad methods and technologies which purportedly allow them to gain future price information. Stock Market Prediction is the act of trying to determine the future value of a company stock or other financial entity traded on an exchange. In a financially volatile market, as the stock market, it is important to have a very precise prediction of a future trend. Because of the financial crisis and scoring profits, it is mandatory to have a secure prediction of the values of the stocks. The successful prediction of stock's future price could yield significant profit.

1.1 Quandl

The premier source for financial, economic, and alternative datasets, serving investment professionals. Quandl's platform is used by over 250,000 people, including analysts from the world's top hedge funds, asset managers and investment banks. Stock values for a range of dates is obtained using Quandl and includes the open, close, high and low values for a given day which will be an important dataset to perform mathematical computation and generate visualizations[6].

1.2 Headlines

The headlines or news will be imported using news APIs into the system and that forms our second dataset that will be used to generate the sentiment of the people towards the company. Sentiment Analysis will be performed [6].

2. Literature Review

2.1 Financial Sentiment Analysis Based on Machine Learning

The aforementioned paper attempted to figure out the best

approach to perform sentiment analysis in aid of predicting stock market movement, Naive Bayes and SVM chief among them. The proposed system benefitted from this paper by observing the use of sentiment analysis in a stock market context environment as well as the efficacy of SVM over Naive Bayes as a text classifier. It assisted in the development of the premise of the proposed system as well as a tentative implementation of basic sentiment analysis [1].

2.2 Stock Prediction using Twitter Sentiment Analysis

The proposed system consists of 2 approaches, that is to say, data mining and sentiment analysis. This particular article/paper helped provide insight into the process of how a sentiment can be mined from textual data. It also assisted the proposed system in learning the kind of vocabulary as is entailed when dealing with the stock market [2].

2.3 Stock Market Prediction using Data Mining

The proposed system consists of 2 approaches, that is to say, data mining and sentiment analysis. This particular article/paper helped provide insight into the process of how conventional methods of data mining go about predicting stock market movement as well as assisted the proposed system in learning where information extracted using sentiment analysis could be helpful in aid of prediction accuracy to the traditional methods [3].

2.4 Prediction of Stock Market using Data Mining and Artificial Intelligence

This paper was studied so as to provide a more, well-rounded context as to the techniques related to data mining used in the stock market. It helped the proposed system ascertain sentiment analysis as the better companion to the traditional data mining approach instead of employing an artificial agent. It was noted, however, that the agent was highly effective in shaving off the delays faced by human operators in placing calls [4].

2.5 Stock Market Prediction using Artificial Neural Networks

As was the case with the previous paper, this paper was studied so as to provide a better-rounded context as to the

Volume 7 Issue 5, May 2018

www.ijsr.net

Licensed Under Creative Commons Attribution CC BY

techniques related to data mining used in the stock market. It helped the proposed system ascertain sentiment analysis as the better companion to the traditional data mining approach instead of employing a neural network in cases which called for the supervised approach. It was noted however that a neural network worked extremely well in situations which called for the unsupervised approach [5].

3. Methodologies

3.1 Datasets

There are 2 types of data sets included in the proposed system. The data sets used for regression models would be imported from online portals such as Quandl. These data sets have comprehensive list of attributes relating to stock prices and are available for download freely in .csv file formats. The reason why we have chosen this data set is because it provides all the information that is required to predict stock values in a simple .csv format which makes it easier to import in the algorithm. The second set of data is for Sentiment Analysis. This data set consists of news headlines/tweets which are imported from online APIs to predict the overall sentiment related to the company stock. Both these data sets are then individually divided into two sets for training and testing purposes.

3.2 Algorithm

On the question of whether to utilize a regression, classification, clustering or association algorithm it is clear that the nature of proposed dataset, association algorithm is out of question right out the gate. Between classification and clustering again it is evident from the supervised nature of the problem, classification is better suited for the problem. That being said, both regression and clustering both have merits to them with the dividing factor being the propensity for continuous values in regression and discrete ones in classification. Given the necessity of classifying any information extracted from the data as one of three classes while also accounting for the continuous nature of the dataset, it is prudent to utilize the regression approach to benefit from the best of both worlds. It is derived from logistic regression but boasts accuracy comparable to those of the more complex algorithms out there.

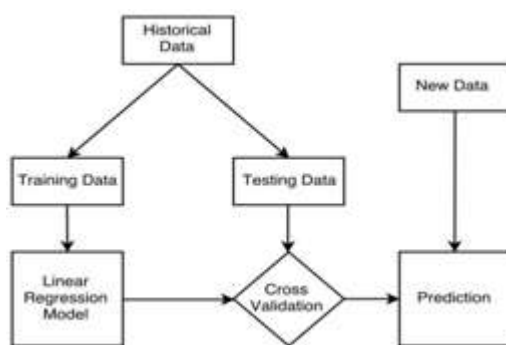


Figure: Flow of Proposed System

3.3 Result



Figure: Conventional Data mining Approach

1096.87153289	1099.76473861	1094.98069756	1099.27272643	1104.38148942	
1119.57788639	1135.70934976	1128.87598516	1122.42524571	1119.51186848	
1117.73591377	1114.77054857	1108.44879439	1103.58702851	1101.45697075	
1119.87497657	1138.93054115	1144.7910756	1159.79697672	1144.04668039	
1163.3338836	1160.20254683	1162.31195149	1179.88001992	1180.35446632	
1190.64236545	1187.34187464	1194.91980374	1214.54089542	1220.48333986	
1225.30251269	1235.08381042	1241.16104638	1238.60155195	1229.11237566	
1235.23496586	1235.43905846	1166.49701535	1118.369728	1127.7702521	
1101.87787903	1047.89697002	1087.7721063	1101.3503034	1101.08765603	
1119.46121984	1140.030137	1143.34448638	1150.92899626	1140.58202108	
1159.15936908	1178.44187896	1184.44122356	1146.90107905	1152.37663579	
1117.94080205	1130.43969847	1142.19738275	1150.65976964	1149.44843142	
1179.27953661	1211.25419825	1216.78518057	1189.02382875	1139.07245956	
1200.98270424	1184.54035117	1148.91347072	1144.4866438	1095.924972121735234 60	
Adj. Close	HL_PCT	PCT_change	Adj. Volume	Volume	
Date					
2017-10-25	991.46	0.299558	0.528225	136042.0	1084.43
2017-10-26	991.42	1.522059	-0.704080	1827682.0	1055.41
2017-10-27	1038.67	2.897443	0.259944	5139945.0	1005.60
2017-10-30	1039.13	0.464515	0.365751	2245352.0	1043.43

Figure: Data mining Approach for Stock market analysis

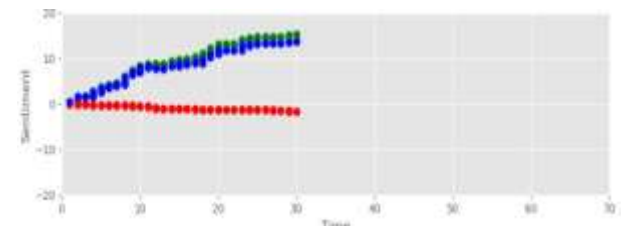


Figure: Sentiment Analysis Approach for predicting stocks

4. Conclusion

The proposed algorithm worked on the views/ opinions of their viewers on the shares. Stock Market is such a field where views of the users matter. The views of the experts affect a lot to the traders who want to enter into the market. The unsupervised and supervised learning depend methods help to find the results in a better way. The combinational study is done to get better accuracy. Further, optimizations can be done in sequence to get improved results.

5. Future Scope

Going forward, the system would heavily benefit from fine-tuning the sentiment target to improve accuracy and value of sentiment analysis to the process. In addition, it would be highly beneficial to improve upon the historical dataset's availability to smooth over even Quandl's minor network/key issues to minimize data unavailability even more. Efforts should be made to incorporate alternative dataset sources to test out their feasibility, for example, Google/Yahoo Finance News, instead of Twitter.

References

- [1] J. Bollen and H. Mao, "Twitter mood as a stock market predictor". IEEE Computer, 44(10):91–94.
- [2] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines." ACM Transactions on Intelligent Systems and Technology, 2:27:1–27:27, 2011.
- [4] G. P. Gang Leng and T. M. McGinnity, "An on-line algorithm for creating self-organizing fuzzy neural networks." Neural Networks, 17(10):1477–1493.
- [5] Lapedes and R. Farber, "Nonlinear signal processing using neural network: Prediction and system modeling." In Los Alamos National Lab Technical Report.
- [6] E. Stefano Baccianella and F. Sebastiani, "Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining." In LREC. LREC.
- [7] <https://www.quandl.com>