International Journal of Pure and Applied Mathematics
Volume 115 No. 6 2017, 521-526
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

USE OF MACHINE LEARNING ALGORITHMS AND TWITTER SENTIMENT ANALYSIS FOR STOCK MARKET PREDICTION

Rohan Pimprikar[1], S. Ramachandran[2], K. Senthilkumar[3]
[1,2]Computer Science & Engineering, SRM University
[1]rohanp965@gmail.com, [2]s.rama2050@gmail.com
[3]Associate Professor (Senior Grade), SRM University

Abstract: This paper experiments with machine learning algorithms and twitter sentiment analysis to evaluate the most accurate algorithm to predict stock market prices. The prediction of stock markets is regarded as a challenging task in financial time series prediction given how fluctuating, volatile and dynamic stock markets are. To aid in dealing with the fluctuations, classifying the sentiment of Twitter data, which oftentimes has been used in finding predictions in a variety of domains, with varying degrees of success according to the particular i. problem statement in question, has been incorporated. ii. However, the use of sentiment classification to predict stock market variables is still challenging but this paper wishes to leverage the current research being done in the scope to improve the accuracy of the predictions. Why sentiment classification will be a huge boost here is because it an exemplary example of moods and psychological states of people at a micro level. Stock markets are heavily sentiment driven and often the panic that precedes a crash or bursting of a bubble, or the excitement in the general public sphere regarding a new technology shape the way the stock price evolves, and social media is often the first to display these trends.

1. Introduction

In a world driven by monetary ambitions, the upside that is potentially offered means that predicting the stock market has established itself on the pinnacle of the areas of finance and engineering. So much capital is channeled through stock trade, it comes as no surprise that the stock market is seen as not just an investment outlet but also a source of income. Additionally, it comes bundled with i. the complexity of proving whether the financial market is ii. predictable or not. There is little that researchers agree iii. upon when it comes to hypotheses that state that attempting to predict the stock market is futile given the variable nature of the same. Hence, still a huge computational war is being waged to prove that the stock market is predictable.

The boom of the Internet and hardware technology means that getting a workstation and accessing the data required is no big ask any more in the current world. And social media, especially a microblogging platform such as Twitter, acts as a constant source of information at a micro level, meaning every individual can have their voice heard rather than an aggregation of thoughts coming out. This can be leveraged to gauge the market sentiment that will be intertwined with recent developments in machine learning prediction algorithms and models, and will together paint a picture of the stock market.

Linear Regression: Linear regression is used to determine the extent to which there is a linear relationship between a dependent variable and one or more independent variables. Before attempting to fit a linear model to observed data, the one modelling should first ascertain whether or not there exists a relationship between the two variables. This however is not an implication that one variable *causes* the other, as causation is not correlation, but instead it conveys that there is a strong relationship between the two variables A scatter-plot is useful as it aids to gauge the degree of association between the two variables. Should there not exist a relationship between the two variables, fitting a linear regression model to the data probably will be an exercise in futility. Scatter plots hinge around the correlation coefficient, which is a measure between -1 to +1 that conveys how strong the relationship between the variables is.

For a linear regression line with the general equation $\boldsymbol{Y = a + bX}$, $\boldsymbol{X}$ is the independent variable while $\boldsymbol{Y}$ is the dependent variable. The slope and the intercept are given as **b** and **a** respectively. **[1]**

Support Vector Machines: Support Vector Machines revolve around the premise of decision planes that segregate space into decision boundaries, based on the different class memberships the space can be split into.

For a Regression SVM:
Y=f(x) + noise.
The important task now is to fit f such that it can bring in fresh cases that the SVM has not yet encountered. This is done by training the SVM model via classification. **[2]**

Multilayer Perceptron Neural Network: Among the two neural network models we have used, multilayer perceptron is a feed forward

artificial neural network model i.e. where the connections between the units don't form a cycle. The main feature of it is that it maps the various sets of data provided as input onto the required outputs. There are a few sheets of nodes similar to a directed-graph, with each of those layers mapped to the next one. Leaving apart the input nodes, each and every node in the multilayer perceptron is actually a neuron (or processing element) with a non-linear activation function.

Input Layer: This layer has three nodes and as no calculations are done here, it passes on the outputs X1, X2 and 1 to the next layer.

Hidden Layer: This layer also has three nodes. The yield of the other two hubs in the Hidden layer is straightforwardly relative on the yields from the Input layer (1, X1, X2) and furthermore with the weights related to the associations. The adjoined figure demonstrates the yield count for one of the concealed hubs (highlighted). Likewise, the yield from other concealed hub can be ascertained. Keep in mind that ***f*** alludes to the activation function. These yields are then provided to the nodes in the Output layer.

Output Layer: This layer has two nodes. It takes contributions from the Hidden layer and performs calculations correspondingly as appeared for the highlighted concealed hub. The qualities ascertained (Y1 and Y2) therefore of these calculations are the yields of the Multi Layer Perceptron which can be additionally removed. **[3]**

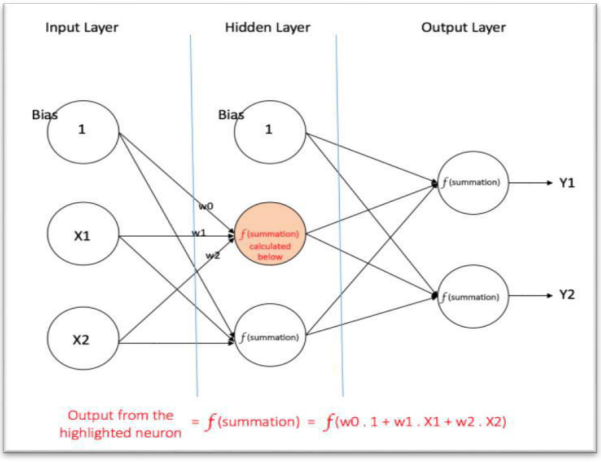

Figure 1. Layers of a Neural Network

Long Short Term Memory: Long Short Term Memory systems– typically just called "LSTMs" – are a unique sort of Recurrent Neural Network, fit for adapting long haul conditions. They are intended to maintain a strategic distance from the long-haul reliance issue. Recollecting data for drawn out stretches of time and at the same time for shorter terms is for all intents and purposes their default conduct.

The way LSTMs work is primarily thanks to the cell state. The cell state, runs straight down the whole chain, with just a couple of direct associations en route. Data courses through it with no noise. The LSTM has the capacity to evacuate or add information to the cell state, managed by structures called Gates. They help information let through. The sigmoid layer outputs are decimals between zero and one, describing how much of each component should be allowed to let through. A value of zero means "let nothing through," while a value of one means "let everything through!" An LSTM has three of these gates, to protect and control the cell state. **[4]**

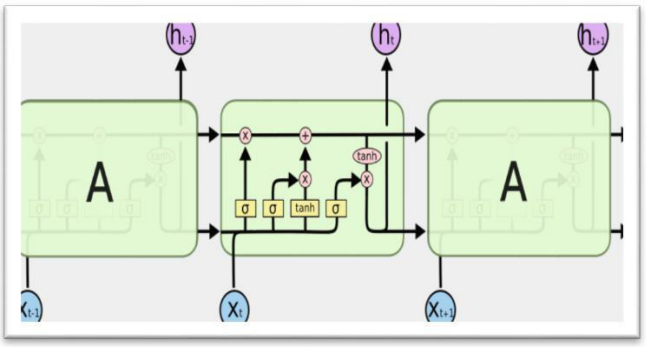

Figure 2. Functioning of inside a cell state

2. Sentiment Analysis

Sentiment Analysis is the process of computationally determining whether a text conveys a positive, negative or neutral sentiment from the user. Opinion mining is also very similar, which derives the opinion or attitude of a speaker. In the marketing field companies use it to develop their strategies, to understand customers' sentiments towards products or brand, how people respond to their product launches and why consumers don't purchase some products. In political field, it helps in keeping track of overall political view, to detect inconsistency and consistency between reactions at the government's perspective. Sentiment analysis is also used to monitor and analyze social phenomena, for the spotting of potentially dangerous situations and determining the general mood of the internet from the various reactions to news articles. **[5]**

3. Literature Survey

i. Existing Research: In **[6]**, the authors analyzed recent advancements in stock market prediction models. By looking at different forecast models, they found that NNs offer the capacity to anticipate showcase headings more precisely than other existing methods. The capacity of NNs to take in nonlinear connections from the preparation input/yield sets empowers them to display non-straight element frameworks, for example, stock markets, all the more definitely. Likewise, by concentrating on a few critical issues in stock markets, they found that that numerous scientists have perceived that subjective elements, for example, political impacts and largescale international events can significantly affect stock costs. It has been looked into that NNs in light of both quantitative and subjective elements are far better than the ones constructed just in light of the quantitative components.

In **[7]**, the authors' key area of concentration was the utilization of sentiment analysis data for machine learning algorithms that enabled them to get most extreme exactness of stock market forecasts for DJIA – 64.10%. For DJIA, their exactness was below the precision that was discovered by Bollen and co-creators in another paper. This could prompt a conclusion that most likely higher expectation rate exhibited by Bollen and co-creators was identified with a little trial (just 19 days).These results could likewise be clarified by different elements. In the first place, it may be the case that data about application of Twitter for DJIA was plainly accessible to the trading society in 2010 and now this examination procedure couldn't reliably beat the market as some of brokers officially utilized it. Somewhat this could affirm effective market speculation. Besides, presumably there was a need to broaden preparing period from 60 days to a while like Bollen and his partners did. Third, they couldn't think about execution straightforwardly on the grounds that restrictive nature of their calculation and further change of the sentimental analyzer was required.

4. Proposed System

Our research revolves around optimizing feature selection in the historical data on stocks which was scraped from the Yahoo Finance website. We have implemented various machine learning models such as Linear Regression, SVM and Neural Networks and tuned them up to the best possible parameters in order to maximize the efficiency. In addition to doing so, we have also incorporated twitter sentiment analysis in opposition to the use of Event Information. Twitter provides a more dynamic, immediate and all-encompassing information about a certain stock which enables us to quantify the information based on positive, negative or neutral reviews. Any news or piece of information about a company that directly impacts its stock price comes almost instantaneously on Twitter before any other news source. Numerous studies have already shown that crucial news information about any major event does influence the stock price.

Feature Selection:
After much iteration, we fixated on these parameters as inputs to the machine learning algorithms:

1. **Adjusted Close**: This is the adjusted closing price of the stock on any given day of trading after factoring in any distributions and corporate actions that occurred at any time prior to the next day's open.

2. **High/Low Percentage**: This is the percentage change in highest price and lowest price that the stock experienced on any given trading day.

3. **Percentage Change in stock price**: This is the percentage change in the opening and closing price that the stock experienced on any given trading day.

Parameters used in the **MLP Regressor** neural network:

1. **Number of hidden layers**: 100
2. **Solver function**: "lbfgs" is an optimizer in the family of quasi-Newton methods.
3. **Activation function**: "relu": the rectified linear unit function, returns f(x) =max (0, x).
4. **All the remaining parameters were left default**.

Parameters used in the **LSTM** neural network:

1. **Number of hidden layers**: 100
2. **Solver function**: "rmsprop" is an adaptive learning rate method.
3. **Activation function**: "linear"
4. **Return_sequences = True**: because we want to retain sequences for next iterations.
5. **All the remaining parameters were left default**.

Sentiment Analysis

Here is how our sentiment classifier is created:

1. We have used a Movies Reviews dataset in which reviews have already been labelled as positive or negative.

2. Positive and negative features are extracted from each positive and negative review respectively.

3. Training data now consists of labelled positive and negative features. This data is trained on a Naive Bayes Classifier.

Naïve Bayes classifiers have a place with a group of basic likelihoods based classifiers which apply Bayes' hypothesis with solid assumptions between the components. They are exceedingly adaptable, requiring the quantity of parameters lineary relative to the quantity of factors (features/predictors) in a learning issue. Maximum-likelihood training was performed on this classifier by assessing a shut-shape expression which takes linear time, as opposed to utilizing costly iterative guess as utilized by numerous different sorts of classifiers. **[8]**

We then quantify the polarity of the tweets between -1 and 1 which is further classified as follows:

Polarity = 0: Neutral Sentiment

Polarity < 0: Negative Sentiment

Polarity > 0: Positive Sentiment

5. Conclusion

The Linear Regression algorithm's accuracy was approximately 82%.The Support Vector Machine algorithm's accuracy was approximately 60%.While, the Multi Layer Perceptron- Regressor neural network's RMSE values were roughly under 0.3 for training and testing data used.

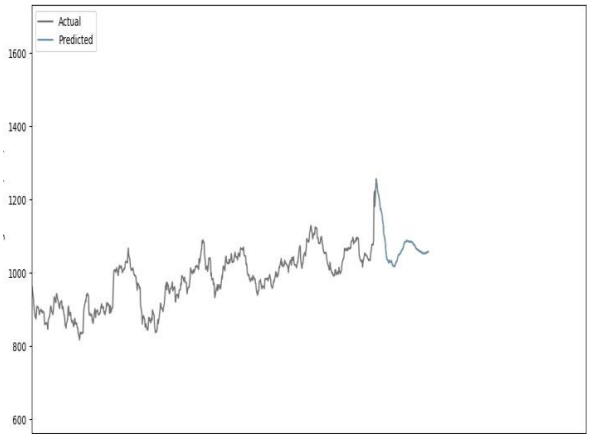

Figure 3 . Result of MLP Neural Network

The LSTM's point by point prediction values were the closest to the actual values. We also observed that increasing the number of hidden layers and the number of input dimensions was computationally expensive and didn't make a significant impact on the accuracy. The resulting diagram is as follows:

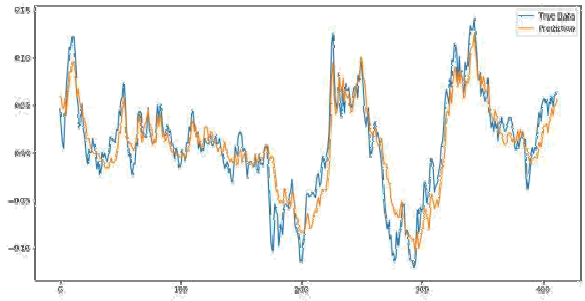

Figure 4. Result of LSTM Neural Network

After implementing Twitter sentiment analysis on stocks, we concluded that it is only influential when certain polarizing news about a company is floating around in the media sources. The analysis can only be considered creditable if there is an extreme polarizing sentiment such as more than 80% tweets are showing a positive sentiment about the stock, then it can be concluded with some certainty that the stock price is bound to go up. Otherwise, the neutral sentiment in the tweets quantitatively overshadows the positive and negative sentiment.

References

[1] David A. Freedman (2009). Statistical Models: Theory and Practice. Cambridge University Press.

[2] Ben-Hur, Asa, Horn, David, Siegelmann, Hava, and Vapnik, Vladimir; vector clustering" (2001) Journal of Machine Learning Research, 2: 125–137.

[3] https://ujjwalkarn.me/2016/08/09/quick-intro-neural-networks/.

[4] Colah, Understanding LSTM Networks, Github 2015.

[5] Turney, Peter (2002). "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews". Proceedings of the Association for Computational Linguistics.

[6] Machine Learning Techniques and Use of Event Information for Stock Market Prediction: A Survey and Evaluation Paul D. Yoo, Maria H. Kim, Tony Jan

Department of Computer Systems Faculty of Information Technology University of Technology, Sydney PO Box 123, Broadway, NSW 2007, Australia.

[7] Machine learning in prediction of stock market indicators based on historical data and data from Twitter sentiment analysis. Alexander Porshnev, Ilya Redkin, Alexey Shevchenko National Research University Higher School of Economics Nizhniy Novgorod, Russia.

[8] Zhang, Harry. The Optimality of Naive Bayes. FLAIRS 2004 conference.

[9] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining." in Lrec, 2010, pp. 1320-1326.

[10] Kim, K. 2003, 'Financial time series forecasting using support vector machines', Neurocomputing, vol. 55, pp. 307- 319.

[11] Twitter Sentiment Classification Using Machine Learning Techniques for Stock Markets Mohammed Qasem, Ruppa Thulasiram, Parimala Thulasiram Department of Computer Science University of Manitoba Winnipeg, Canada.

[12] Ding, T., Fang, V., & Zuo, D. (2013). Stock Market Prediction based on Time Series Data and Market Sentiment. Retrieved from http://murphy.wot.eecs.northwestern.edu/~pzu918/EECS 349/final_dZ uo_tDing_vFang.pdf.

[13] Chen, Ray and Lazer, Marius, "Sentiment Analysis of Twitter Feeds for the Prediction.

[14] Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science.

[15] Kim, K. 2003, 'Financial time series forecasting using support vector machines', Neurocomputing, vol. 55, pp. 307-319.

[16] Kim, K. Hong, T. & Han, I. 1998, 'Knowledge Discovery Process In Internet For Effective Knowledge Creation: Application To Stock Market', Korea Advanced Institute of Science and Technology.

[17] Kim, K. 2004, 'Toward Global Optimization of Case-Based Reasoning Systems for Financial Forecasting', Applied Intelligence, vol. 21, no. 3, pp. 239-249.

[18] Kim, K. and Han, I. 2000, 'Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index', Expert System Appliance, vol. 19.

[19] S. Prasath Sivasubramanian, N. Suganya, "Sentiment Analysis On Micro-blogs", International Innovative Research Journal of Engineering and Technology, vol. 2.