# Use of Machine Learning Algorithms and Twitter Sentiment Analysis for Stock Market Prediction

**Rohan Pimprikar, S. Ramachandran, K. Senthilkumar***

*Associate Professor (Senior Grade), SRM University

Computer Science & Engineering, 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, sentiment classification of Twitter data, which has been successfully applied in finding predictions in a variety of domains, has been incorporated. 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. A key area of said research is the development of linguistic technologies and penetration of social media that helps provide powerful possibilities to investigate users' moods and psychological states of people. 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.***

## I. 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 the complexity of proving whether the financial market is predictable or not. Since there has been no consensus on the validity of Efficient Market Hypothesis (EMH) which states the market is efficient and there is no space for prediction, researchers have strived for proving the predictability of the financial market.

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 acts as a constant source of even information, which has a significant impact on stock markets, the techniques to extract and use information to support decision making have become a critical task. The problem statement of attempting to predict the stock market is as old as anything, and there have been various algorithms that try to achieve the same. In this paper, recent developments in prediction algorithms and models will be introduced and their performance will be compared. In addition, for accurate market sentiments, tweets from the microblogging social network platform Twitter will be considered.

i. **Linear Regression**: Linear regression can be explained as an attempt to model the relationship between two variables by fitting a linear equation to the observed data. Of the two variables, one is the explanatory variable with the other being the dependant one. Before attempting to fit a linear model to observed data, the one modelling should first determine whether or not there is a relationship between the variables of interest. 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 gauge the degree of association between the two variables. If there appears to be no association between the proposed explanatory and dependent variables (i.e., the scatter plot does not indicate any increasing or decreasing trends), then fitting a linear regression model to the data probably will not provide a useful model. A valuable numerical measure of association between two variables is the correlation coefficient, which is a value between -1 and 1 indicating the strength of the association of the observed data for the two variables.

A linear regression line has an equation of the form ***Y* = *a* + *bX***, where ***X*** is the explanatory variable and ***Y*** is the dependent variable. The slope of the line is ***b***, and ***a*** is the intercept (the value of ***y*** when ***x*** = 0). **[1]**

ii. **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.**

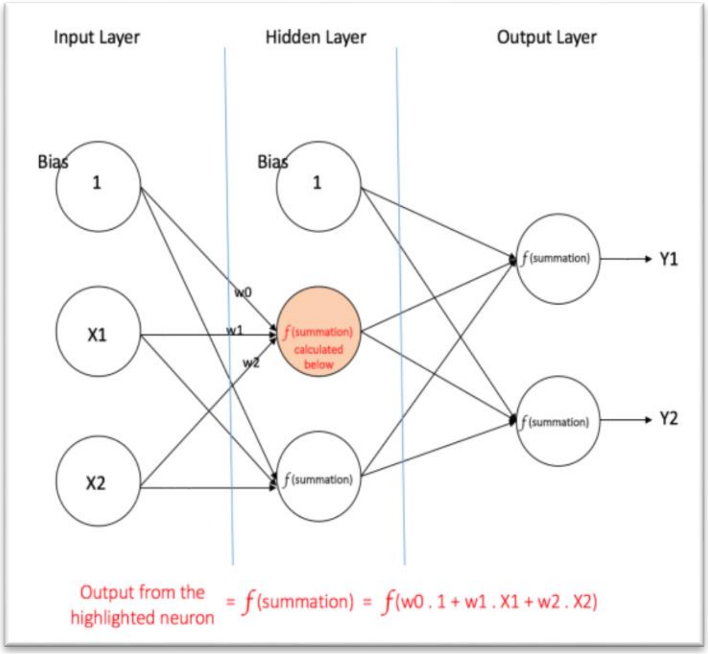
The task is then to find a functional form for f that can correctly predict new cases that the SVM has not been presented with before. This can be achieved by training the SVM model on a sample set, i.e., training set, a process that involves, like classification (see above), the sequential optimization of an error function. **[2]**

iii. **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. The Bias node carries the value 1. The other two nodes take X1 and X2 as input feeds externally (which are numerical values depending upon the input dataset). There is no computation performed in the Input layer, so the outputs from nodes in the Input layer are 1, X1 and X2 respectively, which are fed into the Hidden Layer.

**Hidden Layer:** This layer also has three nodes. The Bias node has an output value of 1. The output of the other two nodes in the Hidden layer is directly proportional on the outputs from the Input layer (1, X1, X2) and also with the weights associated to the connections. The incorporated figure shows the output calculation for one of the hidden nodes (highlighted). Similarly, the output from other hidden node can be calculated. Remember that ***f*** refers to the activation function. These outputs are then supplied to the nodes present in the Output layer.

**Output Layer:** The Output layer has two nodes. It takes inputs from the Hidden layer and performs computations similarly as shown for the highlighted hidden node. The values calculated (Y1 and Y2) as a result of these computations are the outputs of the Multi Layer Perceptron which can be further extracted. **[3]**


Input Layer
Hidden Layer
Output Layer
Bias
1
X1
X2
Bias
1
w0
w1
w2
f(summation) calculated below
f(summation)
f(summation)
f(summation)
Y1
Y2
Output from the highlighted neuron = f(summation) = f(w0 . 1 + w1 . X1 + w2 . X2)

i. **Long Short Term Memory:** Long Short Term Memory networks – usually just called "LSTMs" – are a special kind of Recurrent Neural Network, capable of learning long-term dependencies. They are designed to avoid the long-term dependency problem. Remembering information for long periods of time and simultaneously for shorter durations is practically their default behavior.

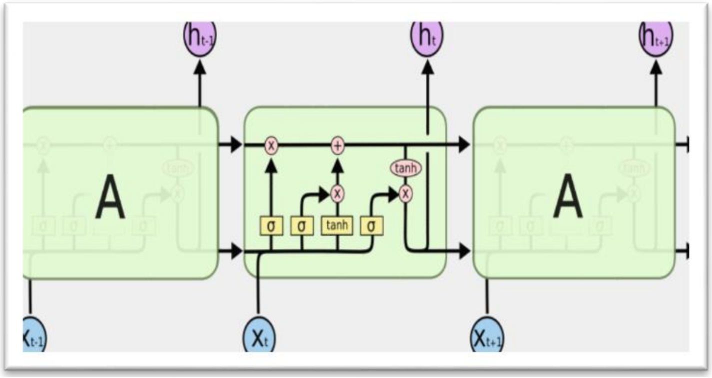
The key to LSTMs is the cell state. The cell state, runs straight down the entire chain, with only a few linear interactions along the way. Information flows through it without any noise.The LSTM does have the ability to remove or add information to the cell state, carefully regulated by structures called gates. Gates help information let through. They are composed out of a sigmoid neural net layer and a point wise multiplication operation. 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.

The primary step in the LSTM is to classify the information which will be thrown away from the cell state. This decision is made by a sigmoid layer called the “forget gate layer.” It looks at $h_{t-1}$ and $x_t$, and outputs a number between 0 and 1 for each number in the cell state $C_{t-1}$. A 1 represents “completely allow” while a 0 represents “completely get rid”

The next step involves deciding what prevalent information will be stored in the cell state. This has two parts. First, a sigmoid layer called the “input gate layer” decides which values we’ll update. Secondly, a tanh layer creates a vector of new candidate values, $C\sim_t$, that could be added to the state.

The next step is to update the old cell state, $C_{t-1}$, into the new cell state $C_t$. The previous steps already stated what to do.We multiply the old state by $f_t$, forgetting the things to remove we decided earlier. Then we add $i_t*C\sim_t$. These are the new candidate values, scaled upon by how much we decided to update each state value.

The final step is to decide the output of the cell state. This output will be a filtered version but based on the cell state. First, the sigmoid layer decides which parts of the cell state shall be referenced to output. Then, we put the cell state through tanh (to limit the values between −1and 1) and multiply it by the output of the sigmoid gate, so that we only output the parts we earlier decided to. **[4]**



**II. 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]**

## III. LITERATURE SURVEY

### i. Existing Research:

In **[6]**, the authors examined recent developments in stock market prediction models. By comparing various prediction models, they found that NNs offer the ability to predict market directions more accurately than other existing techniques. The ability of NNs to learn nonlinear relationships from the training input/output pairs enables them to model non-linear dynamic systems such as stock markets more precisely. In addition, by studying several important issues in stock markets, they found that that many researchers have recognized that qualitative factors such as political effects and international events can have a significant impact on stock prices. It has been reviewed that NNs based on both quantitative and qualitative factors are far superior to the ones based only on the quantitative factors.

In **[7]**, the authors’ key area of focus was the application of sentiment analysis data for machine learning algorithms that allowed them to receive maximum accuracy of stock market predictions for DJIA – 64.10%. For DJIA their accuracy was below the 87.6% accuracy that reported by Bollen and co-authors in another paper. This could lead to a conclusion that probably higher prediction rate demonstrated by Bollen and co-authors was related to a small test period (only 19 days).These results could also be explained by other factors. First, it could be that information about application of Twitter for DJIA become available to trading society in 2010 and now this analysis technique could not consistently beat the market as some of traders already used it. Partially this could confirm efficient market hypothesis. Secondly, probably there was a need to extend training period from 60 days to several months like Bollen and his colleagues did. Third, they could not compare performance directly because proprietary nature their algorithm and further improvement of our sentiment analyzer needed. However, they found out that Support Vector Machine provide a little better prediction accuracy of S&P500 indicator (62.03%), than 51.88% demonstrated by Ding et al."

ii. **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.

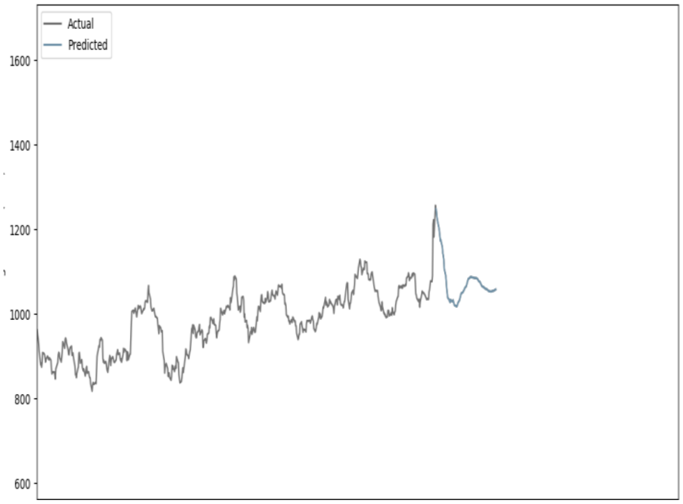
**Naïve Bayes classifiers** belong to a family of simple probability based classifiers which apply Bayes' theorem with strong assumptions between the features. They are highly scalable, requiring the number of parameters lineary proportional to the number of variables (features/predictors) in a learning problem. Maximum-likelihood training was performed on this classifier by evaluating a closed-form expression which takes linear time, rather than using expensive iterative approximation as used by many other types of classifiers.**[8]**

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

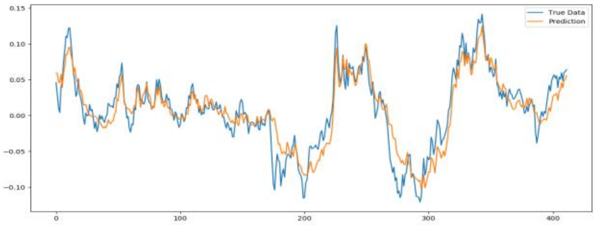
**Polarity = 0: Neutral Sentiment**
**Polarity < 0: Negative Sentiment**
**Polarity > 0: Positive Sentiment**

## IV. 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.



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:



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.

## V. 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; "Support 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 fromhttp://murphy.wot.eecs.northwestern.edu/~pzu918/EECS349/final_dZ uo_tDing_vFang.pdf

[13] Chen, Ray and Lazer, Marius, “Sentiment Analysis of Twitter Feeds for the Prediction of Stock Market Movement,” Stanford, 2013

[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.