BIUST Research and Innovation Symposium 2017 (RDAIS 2017)
Botswana International University of Science and Technology
Palapye, Botswana, 12-14 June 2017

ISSN: 2521-229X
ISBN: 978-99968-0-608-7

Testing the Predictability of the Botswana Stock Exchange: Evidence from Supervised Machine Learning

Kushatha Kelebeng
Department of Computer Science and Information Systems
BIUST
Palapye, Botswana
kushatha.kelebeng@studentmail.biust.ac.bw

Hlomani Hlomani
Department of Computer Science and Information Systems
BIUST
Palapye, Botswana
hlomanihb@biust.ac.bw

***Abstract*—Prediction of the stock market is a vital part of the economy especially for emerging markets in developing countries. There is significant literature on predicting the stock market particularly in developed countries like the US. However, there is need for more research in emerging markets such as developing countries like the Botswana Stock Exchange. This paper aims at evaluating the predictability of the Botswana Stock Exchange using supervised machine learning to specifically assess and test the null hypothesis of the Random Walk Theory. Machine learning is one of the upcoming trends of data mining; hence few machine learning algorithms have been used where their results have been compared using classification evaluation parameters such as Accuracy, Mean Average Error (MAE), Receiver Operating Characteristic Area (ROC), Kappa Statistic, Precision and Recall. Naïve Bayes have been considered the most effective model as it yielded the highest accuracy of 83.3% with the least error margin. The results reject the null hypothesis of the Random walk Theory for Botswana Stock Exchange for the period of January-December 2015, clearly indicating that the Botswana Stock market is predictable using machine learning techniques.**

Keywords*—*Emerging Markets, Machine learning, Random Forest, Naïve Bayes, Support vector machines

I. INTRODUCTION

Several studies have been conducted to predict the future stock price direction. Such studies would mainly use methods in financial analysis and data analysis. In recent years, machine learning has become a widely used method for stock prediction. The most used methods are Artificial Neural Networks and Support Vector Machine [1]. Even though a significant number of researches have been done on predicting the stock price index, most researches that have been done are on the developed markets. However, few researches exist on predicting the direction of stock price in emerging markets, especially in the Botswana Stock Market.

Accurately predicting the market is important because it helps in developing effective market strategies, which in turn, helps investors, mitigate risks and make profit [2]. Though it is very important, it is at the same time very difficult to predict the stock market because it is non-linear and is affected by a number of factors such as political, social and economic.

It is of great interest to study the prediction of the stock market using data from emerging markets such as that of Botswana. Since its establishment in 1989, solidified by the Botswana Stock Exchange Act of 1994, the Botswana Stock Exchange (BSE) has grown tremendously as an emerging market. The BSE is one of Africa's best performing stock exchange averaging 24% aggregate returns in the past decade. The BSE is the third largest stock exchange in terms of capitalization in Southern Africa. It has 35 marked listings and 3 stock indices; the Domestic Company Index, Foreign Company Index and the all Company Index [3]. The objective of this paper is to predict the direction of the BSE using machine learning techniques; Random Forest, Support Vector Machines and Naïve Bayes. The contribution of this paper is in demonstrating that the stock price is predictable using supervised machine learning techniques and in comparing the performance of these techniques on a given set of metrics.

II. RELATED WORK

This section gives a background and a synopsis of previous researches that have been carried out in the areas of stock market predictability, and machine learning techniques (Support vector machines, Random forest and Naïve Bayes). The purpose of the section is to ease the user into the subject matter through definition of important concepts and giving the necessary background discussion to motivate the problem under study.

A. Predictability of the Botswana Market

The random walk is a theory that states that stock prices are independent of each other though they have the same distribution [4]. This theory implies that past stock prices cannot be used to predict future stock prices. Even though a lot of research has been done in predicting the stock market, little research has been done on emerging markets like the Botswana market. The need for research in emerging markets is still not recognized. Radikoko [3] tested the effectiveness of the

ISSN: 2521-229X
ISBN: 978-99968-0-608-7

Botswana market. He was particularly interested in whether the random walk theory applies to the Botswana market. The author used different methods in order to validate his research.

The results presented by the authors showed that the random walk theory does not govern the Botswana stock market which meant that the market is weak-form inefficient. The results also meant that historical prices could be used to predict future stock prices. Similarly, Mollah in 2007 [2] tested the null hypothesis of the random walk theory by testing whether the Botswana Stock Exchange is predictable. The author used data from the Botswana stock exchange from 1989 to 2005. The author found that the Botswana market can be predicted using the Triangulation approach. Their results also nullified the null hypothesis of the random walk theory, suggesting that the market was predictable.

Chiwira and Muyambiri did some research in the same area in 2012 [5]. The research evaluated the efficiency of the Botswana Stock Exchange by assessing the random walk theory using a number of methods. Their research rejected the random walk hypothesis. The authors used BSE data from 2004 to 2008, which constituted monthly and weekly data. They showed that the market could be outperformed since stock prices are not independent of past price changes. The authors concluded that both fundamental and technical analysis can bring positive results. All these studies show that indeed the Botswana market is predictable. They therefore set a baseline for research in this field and also motivates for more research work to be done on predicting the Botswana stock market.

B. Stock Price Behaviour: Theories

The stock market is guided by a number of indicators namely; Technical analysis, Fundamental analysis, efficient market theory and the random walk theory. Technical analysis is a methodology for forecasting the direction of security prices through the study of past market data. Technical analysis examines statistics generated by the market using past prices and volumes in order to evaluate securities. Technical Analysis is the study of past prices in order to predict future prices, patterns and trends. Technical analysis studies the patterns on the market, the supply and the demand of stock shares [10]. There has been a lot of debate over whether technical analysis can actually predict the markets. Previous studies [6], [7], [8] examined the predictability of the stock market through the use of technical analysis. These studies failed to provide evidence of the usefulness of technical analysis since they had mixed conclusions [9].

Fundamental Analysis on the other hand, is the study of financial information like company earnings and value of assets. With fundamental analysis, everything is studied from financial conditions to the industrial conditions and even the management of the company [11]. Fundamental analysis evaluates a share through measuring its value by examining factors such as financial, economic, quantitative and qualitative. According to Suresh [12], fundamental analysis is actually the study of factors that affect the economy and companies. Much like technical analysis (when applied to stock markets), the goal of fundamental analysis is to predict the future stock price movement and thus making profit from it.

The random walk theory implies that the stock price fluctuates and its fluctuations are independent and may be described by a random process like tossing a coin, for example. It states that the current stock market price is independent and unrelated to previous market price patterns. The theory implies that stock price changes do not have memory, cannot be predicted basing on past history of the behaviour of the stock. However, the actual price of the stock may change in response to new information. The value of a stock is determined by fundamental analysis of the future earnings performance of the company. As new information becomes available investors may revise their estimates of expected future earnings and those revisions may affect the estimated value of the stock [13]. Furthermore, Poshakwale [14] states that, it refers to the fact that price changes are independent of each other. Tomorrow's price change and tomorrow's price cannot be predicted by looking at today's price change.

The Efficient Market Hypothesis (EMH) is a market theory that evolved from a 1960's Ph.D. dissertation by Eugene Fama [4]. The efficient market hypothesis states that, at any given time and in a liquid market, security prices fully reflect all available information. The EMH exists in various degrees: weak, semi-strong and strong, which addresses the inclusion of non-public information in market prices. This theory contends that since markets are efficient and current prices reflect all information, attempts to outperform the market are essentially a game of chance rather than one of skill. The weak form of EMH assumes that current stock prices fully reflect all currently available security market information. It contends that past price and volume data have no relationship with the future direction of security prices. It concludes that excess returns cannot be achieved using technical analysis.

The semi-strong form of EMH assumes that current stock prices adjust rapidly to the release of all new public information. It contends that security prices have factored in available market and non-market public information. It concludes that excess returns cannot be achieved using fundamental analysis. The strong form of EMH assumes that current stock prices fully reflect all public and private information. It contends that market, non-market and inside information is all factored into security prices and that no one has monopolistic access to relevant information. It assumes a perfect market and concludes that excess returns are impossible to achieve consistently [15].

C. Random Forests

Random forests are powerful learning algorithms that are used in classification tasks. Training a random forest generates a myriad of decision trees which are then classified based on the results obtained from those decision trees [16]. The mode of the targeted outputs from each decision tree is the output of the forest. Random Forest average out the multitude of the decision tree whereas other trees tend to overfit data because of low bias and high variance. They use random samples of the training data to generate decision trees. This reduces the variance in the overall model thus improving performance and also controlling overfitting. In classification, tree nodes

ISSN: 2521-229X
ISBN: 978-99968-0-608-7

represent features where important features are high up in the tree. Class labels are represented as leaves. The importance of a feature is determined by a Gini impurity where, the lesser the decrease in accuracy by randomly permuting the values of the feature, the less important the feature is [17].

Random Forest classifiers are more accurate and robust to noise and outliers than single classifiers. They run effectively on larger datasets and can handle thousands of input variables without variable deletion. They give an estimate of what variables are most important in the classification; it generates an internal unbiased estimate of the generalization error. Proximities between pairs of cases that can be used in locating outliers are computed using Random Forest. They contain a combination of classifiers where each classifier contributes with a vote for most frequent class to be assigned to the input vector [17]. Random Forest classifier is different to traditional classification trees since it is a combination of many classifiers and thus has special characteristics; it is an ensemble of classification algorithms, which use trees as their base classifiers. They can use some of the data more than once when training and the other data may not be used at all. This makes it achieve more stability, classification accuracy, and makes it more robust when facing variations in input data. Since they are based on bagging, they are not sensitive to noise and over training [18].

D. Naïve Bayes

The Naïve Bayes Classifiers are based on statistics. They are based on the Bayes theorem. The Naïve Bayes classifiers can predict class membership probabilities (i.e. the probability that a given sample belongs to a certain class). The classifier is based on an assumption called the class conditional independence, which assumes that the effect of an attribute value on a given class is independent of the values of an attribute [19]. Naïve Bayes classifiers are linear classifiers that are simple yet very efficient. The assumption is that, features in a dataset are mutually dependent thus the adjective, naïve. Though the independence assumption is normally violated, Naïve Bayes classifiers perform very well even under unrealistic assumptions. Naïve Bayes classifiers can even outperform the most powerful alternatives especially for small size samples. They can perform poorly when the independence assumptions are violated and when they are dealing with non-linear classification problems [20].

The naïve Bayes classifier is mainly used in machine learning models when it comes to the issue of document classification. The Naïve Bayes classifier is linear, meaning that a line on the vector model can separate all of its data samples. The word 'naive' means that all features within the dataset are independent of one another. This naive assumption states that the value of any particular feature is unrelated to and does not depend on the value or presence of any/all other features, given the label [21]. The success of Naive Bayes in the presence of feature dependencies can be explained by the fact that optimality in terms of the classification error is not in any way related to the quality of the fit to a probability distribution such as the appropriateness of the independence assumption. An optimal classifier is obtained when the actual and estimated distributions agree on the most probable class [22]. The probability model for a classifier is a conditional model

$$P(cj|d) = p(cj|x1, x2, x3, x4....xn)\, P(cj) \quad (1)$$

In equation 1, x1, x2, x3, x4…xn are features or words of a document. Cj is a set of classes used in classification, p (cj|d) is a conditional probability and P (cj) is the prior probability of class c. If a feature has large values or if the number of features is large then it is difficult to calculate the probability, the model will then have to be manipulated by changing parameters and filtering some features. According to Narayanan et al. [23], the Naive Bayes includes a simplifying conditional independence assumption. For example, if there is a given class (positive and negative), the two words are conditionally independent of one another. The accuracy of text classification is not affected by this assumption [23].

E. Support Vector Machines

Vapnik [24] originally proposed the Support Vector Machine (SVM) classifier in 1995. It finds a maximal margin separating hyperplanes between two classes of data. There are non-linear extensions of the SVM. The SVM classifier is used to classify both linear and non-linear data using a margin. It is a margin based classifier which works by selecting the maximum margin. It separates the classes with a surface that maximizes the margin. Support vectors are the data points near to the margin. The hyperplanes are the decision boundaries. In SVM, the data is trained using the known labeled classes and then a model is built. Elements of training data generated during SVM learning module are support vectors. The model and the vectors are used to classify the data [24].

Intuitively, a good separation is achieved by the learned hyperplane that has the largest distance to the nearest training data points of any class (so-called functional margin), since in general, the larger the margin the lower the generalization error of the classifier. Support Vector Machines provide an easy to use interface to the world of SVMs. The major drawback with SVM is that it scales badly with the size of the data due to the kernel transformation and the quadratic optimization. Therefore, choosing correct kernel parameters is important for obtaining good results. Support Vector Machines are binary classifiers though they can be used for multi-class classification by two methods being the: one vs. all and the one vs. one. For the one vs. one, for k classes, $k\,(k-1)/2$ binary classifiers are trained. Each classifier is trained with a pair of classes from the original training set and a hyperplane is learned between classes. A voting scheme is used to make predictions. In one vs. all, only a single classifier is trained in a class by labeling examples as positive and negative. This requires the base classifiers to produce a real score for its decision other than just a label. Predictions are then made using these scores [25].

III. EXPERIMENTATION

The proposed method in this paper uses general public mood as an input attribute to the prediction model. The output of the model is one of the four defined classes from the Google profile of Moods state; kind, vital, alert and happy. The model uses three machine learning algorithms and compares their

ISSN: 2521-229X
ISBN: 978-99968-0-608-7

results. The mood was normalised between [+1, 0, -1] because moods can have positive, neutral and negative values.

A. Scope

The data that was used in this paper spread over twelve months, from January, 2015 to December, 2015. The sentiments were collected from Facebook using Discover Text crawler. This was data from an official Facebook page of a company listed under the domestic index on Botswana Stock Exchange. The historical stock prices were collected from Botswana stock Exchange. It was opening and closing stock price of that particular day.

B. Attributes Used

1. Historical Stock Price Data

The first attribute used as an input for the model is the historical closing index of Botswana Stock Exchange. The historical data was not made part of the model directly.

2. Public Mood

The performance of the stock market can be influenced by the moods and sentiments of investors. The stock market performance is driven by collective sentiments. Through social media we can be able to achieve this. In this paper, Facebook is used as a source of public sentiment.

C. Experimentation

1. Data Collection

Data was collected and pre-processed so that it can be in a form that is acceptable as input to the model. The Facebook posts and comments for each day were gathered and categorized as kind, vital, alert or happy. About 3000 instances were collected for twelve months (from January to December 2015). Out of these instances, 60% was used in training and the remaining 40% were included in the test data set.

2. Implementing Machine Learning Algorithms

All three machine-learning algorithms were separately applied on the data sets and measures of their accuracy were obtained on two main aspects: 10 Cross Validation and 60% Data Split. For the SVM algorithm the SMO implementation in WEKA was used. Accuracy in this case considers the overall effectiveness of the classifier [26].

IV. RESULTS AND DISCUSSIONS

1. Results Presentation

Table 1 depicts the accuracy results of the classifiers. The SVM classifier gave 83.3% results when demonstrated on cross validation and 80.0% when demonstrated on data split. Applying Naïve Bayes algorithm on cross validation and Data split gave the same results as SVM. The algorithm produced 83.3% accuracy on cross validation but 80% on the data split. The Random forest algorithm did not perform well with 66.0% on cross validation and 20.0% accuracy on data split.

TABLE I. RESULTS FOR THE ACCURACY METRIC

Algorithm	**Accuracy**	
	10 Cross Validation	***60% Data Split***
Random Forest	66.0%	20.0%
Naïve Bayes	83.3%	80.0%
SVM	83.3%	80.0%

Table II on the other hand, depicts precision results. Precision is the measure of results relevancy. It is the agreement between the data labels and positive labels given by the classifier. Good precision means less false negatives [27] as defined in Formula 2.

$$P = \frac{Tp}{Tp+Fp} \quad (2)$$

Precision is the number of true positives over the number of true positives plus the number of false positives [27].

TABLE II. RESULTS FOR THE PRECISION METRIC

Algorithm	**Precision**	
	10 Cross Validation	***60% Data Split***
Random Forest	0.708	0.100
Naïve Bayes	0.900	0.867
SVM	0.900	0.867

From Table II, it is clear that the Naïve Bayes and SVM classifiers perform better than the Random Forest classifier in terms of precision in both test methods. The Naïve Bayes model gave 0.900 when demonstrated on cross validation and 0.867 when demonstrated on data split. The SVM algorithm produced 0.900 precision on cross validation but 0.867 on the data split. The Random forest algorithm produced 0.708 on cross validation and 0.100 precision on data split.

Recall is the measure of how many truly relevant results are returned. It is the effectiveness of the classifier to identify positive labels. Good recall means less false negatives [27].

$$R = \frac{Tp}{Tp+Fn} \quad (3)$$

Recall as depicted in Formula 3, is the number of true positives over the number of true positives plus the number of true negatives [27].

ISSN: 2521-229X
ISBN: 978-99968-0-608-7

TABLE III. RESULTS FOR THE RECALL METRIC

Algorithm	Recall	
	10 Cross Validation	*60% Data Split*
Random Forest	0.667	0.200
Naïve Bayes	0.833	0.800
SVM	0.833	0.800

From Table III, Naïve Bayes and SVM classifiers perform better than the Random Forest classifier in terms of recall in both test methods. The Naïve Bayes model gave 0.833 when demonstrated on cross validation and 0.800 when demonstrated on data split. The SVM algorithm produced 0.833 on cross validation but 0.800 on the data split. The Random forest algorithm produced 0.667 on cross validation and 0.200 recall on data split.

The ROC is a graphical plot for organising, visualising and selecting classifiers based on performance. It is created by plotting true positive rate against false positive rate. According to literature an optimal classifier would be more than 0.5 and not more than 1.00 [26]

TABLE IV. RESULTS FOR THE ROC METRIC

Algorithm	ROC	
	10 Cross Validation	*60% Data Split*
Random Forest	0.907	0.867
Naïve Bayes	1.00	0.867
SVM	0.935	0.867

From Table IV, the Naïve Bayes model gave 1.00 when demonstrated on cross validation and 0.867 when demonstrated on data split. The SVM algorithm produced 0.935 ROC metric on cross validation but 0.867 on the data split. The Random forest algorithm produced 0.907 on cross validation and 0.867 on data split. What is more interesting about this results is that for 60% data split all the three algorithms had the same results. Kappa compares observed accuracy with expected accuracy. Kappa of 1 indicates a perfect agreement and of 0 indicates agreement by chance [27]. The formula for calculating Kappa statistic is depicted in Formula 4.

$$K=\frac{P(A)-P(E)}{1-P(E)} \quad (4)$$

TABLE V. RESULTS FOR THE KAPPA METRIC

Algorithm	Kappa	
	10 Cross Validation	*60% Data Split*
Random Forest	0.556	0.1304
Naïve Bayes	0.788	0.6875
SVM	0.788	0.6875

From Table V, the Naïve Bayes model gave 0.778 when demonstrated on cross validation and 0.6875 when demonstrated on data split. The SVM algorithm produced 0.788 kappa on cross validation but 0.6875 on the data split. The Random forest algorithm produced 0.556 on cross validation and 0.1304 accuracy on data split.

Mean Absolute Error (MAE) is how close predictions are to the eventual outcomes as depicted in Formula 5 [27].

Average of absolute errors $|e_1| = |f_1 - y_1|$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|f_1 - y_1| = \frac{1}{n}\sum_{i=1}^{n}|e_i| \quad (5)$$

TABLE VI. RESULTS FOR THE MAE METRIC

Algorithm	Mean Absolute Error	
	10 Cross Validation	*60% Data Split*
Random Forest	0.2267	0.294
Naïve Bayes	0.0813	0.1501
SVM	0.2639	0.2667

From Table VI, the Naïve Bayes model gave 0.0813 when demonstrated on cross validation and 0.1501 when demonstrated on data split. The SVM algorithm produced 0.2639 on cross validation but 0.2667 on the data split. The Random forest algorithm produced 0.2267 on cross validation and 0.294 accuracy on data split.

2. *Performance Comparison of the Algorithms*

The results of all three algorithms over cross validation and data split are compared on Table VII. It is evident from the comparison table that the SVM and Naïve Bayes performed best on cross validation while the Random Forest algorithm did not do well on both cross validation and data split. Conventionally, the verification of the model was done using the Mean Absolute Error (MAE). Therefore, Naïve Bayes seems to be more efficient in predicting the market performance because it had less error margins. While verifying the model's variants, Naïve Bayes outperforms the remaining two algorithms.

TABLE VII. COMPARISON OF MACHINE LEARNING ALGORITHMS

Algorithm	Accuracy	
	10 Cross-Validation	*60% Data Split*
Random Forest	66.0%	20.0%
Naïve Bayes	83.3%	80.0%
SVM	83.3%	80.0%

ISSN: 2521-229X
ISBN: 978-99968-0-608-7

V. CONCLUSION

In this paper, we apply three different algorithms of machine learning to forecast movement direction of Botswana Stock Index from Botswana stock market. In this paper we presented a machine learning methodology for stock price prediction. Two algorithms, being the SVM and the Naïve Bayes produce good prediction with hit rate more than 80%. The SVM and Naïve Bayes outperformed the Random Forest algorithm. Overall, the results of this study confirm that machine-learning techniques are capable of predicting the stock market performance. Botswana Stock Market does follow a behaviour that can be predicted using machine learning techniques. The Naïve Bayes algorithm of machine learning predicted 83.3% correct market performance. Even with the lack of resources and unavailability of data for the market, the model was able to predict the performance of the model to a good extent showing that BSE can be predicted using machine learning techniques. Therefore, SVM and Naïve Bayes are recommended for forecasters of stock index movement and the better model, SVM, is more preferred since it had less error margins. We have presented a machine learning approach toward predicting a company's financial performance using Facebook posts and comments that are related to them from Facebook. Three different classification algorithms (Random forest, Naive Bays and SVM) are used to find the best model for prediction. Our experiment shows that with an accuracy of 83.3% Naïve Bayes can predict whether a company will over-perform or under perform.

VI. RECOMMENDATION

This research can be extended by including more companies to check if the same prediction accuracy will still apply. Twitter can also be used for the same analysis where companies are more active in Twitter than Facebook. Other media forums other than Twitter and Facebook can be used to check the optimal accuracy since those forums may reflect the companies much better. It has been shown in this paper the ability of the SVM, Random Forest and the Naïve Bayes to predict the closing stock price of a selected company. As a result some recommendations can be made that there is need to evaluate the performance of these algorithms over a larger set of data; the need of identifying other variables apart from social media data which also might be used in determining the stock prices.

REFERENCES

[1] V. Shahpazov, B. V. Velev, and A. L. Doukovska, "Design and application of ANN for predicting the values of indexes on the Bulgarian stock market," Signal processing symposium, 2013.

[2] S.A. Mollah, "Testing the Weak Form market Efficiency in Emerging Markets: Evidence from Botswana Stock Exchange," International Journal Of Theoretical and Applied Finance, vol. 10, no. 6, pp. 1077-1094, 2007.

[3] I. Radikoko, "Testing the Random Walk behaviour of Botswana's Equity Returns," Journal Of Business Theory and Practice, p. 84, 2014.

[4] E. F. Fama and M. E. Blume, "Filter rules and stock market trading," Journal of Business, vol. 39, no. 1, pp. 226-241, 1966.

[5] O. Chiwira and B. Muyambira, "A Test of Weak Form Efficiency for Botswana Stock Exchange," British Journal of Economics, Management and Trade, vol. 2, no. 2, pp. 83-91, 2012.

[6] W. Brock, J. Lakonishok, and B. Lebaron, "Simple Technical Trading rules and the Stochastic Properties of Stock returns," Journal of Finance, vol. 47, no. 5, pp. 1731-1764, 1992.

[7] M. C. Jensen and G. A. Benington, "Random walks and technical theories: some additional evidence," Journal of Finance, vol. 2, no. 25, pp. 469-482, 1970.

[8] J. Fang, Y. Qin, and B. Jacobsen, "Technical market indicators: An overview," Journal of Behavioural and Experimental Science, vol. 4, pp. 25-56, 2014.

[9] H. Achelis, Technical Analysis from A-Z.: Vision Books, 2000.

[10] B. G. Malkiel, "The Efficient Market Hypothesis and its Critics, "Journal of Economic perspectives, vol. 17, no. 1, pp. 59-82, 2003.

[11] J. Van Horne and G. G. Parker, "The Random Walk Theory: Empirical Test," Financial Analysis Journal, 1967.

[12] A S Suresh, "A study on fundamental and technical analysis," International Journal of marketing, financial services and management research, vol. 2, no. 5, May 2013.

[13] J Grossman S and E Stiglitz J, "On the impossibilities of informationally efficient markets," The American economic review, vol. 10, no. 3, pp. 393-408, 1980.

[14] S. Poshakwale, "Evidence on Weak Form Efficiency and Day of the Week Effect in the Indiana Stock Market," Finance India, no. 3, pp. 605-616, 1996.

[15] A. Liaw and M. Wiener, "classification and regression by random forest," vol. 2, no. 3, December 2002.

[16] V. Svetnik et al., "Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modelling," Journal of Chemical Information and Computer Sciences, vol. 43, no. 6, pp. 1947-1958, November 2003.

[17] V. F. Rondiguez-Galiano, B. Ghimere, J. Rogan, M. Chica-Olmo, and J. P. Rigol-Sanchez, "An Assessment of the effectiveness of a random forest classification for land classification," Journal of photogrammetry and remote sensing, pp. 93-104, 2012.

[18] K. M. Leung, "Naive Bayes Classifier," Polytechnic University, Department of Computer Science, 2007.

[19] S. Raschka, Naive Bayes and Text Classification: Introduction and Theory, 2014.

[20] S.C. Gangireddy, "Supervised Learning for Multi-Domain Text Class," San Jose State University, San Jose, 2016.

[21] I. Rish, An Empirical Study of the Naive Bayes Classifier, Unpublished

[22] V. Narayanan, L. Arora, and A. Bhatia, "Fast and Accurate Sentiment Classification using an Enhanced Naive Bayes Model," India Institute of Technology, Varanasi, India.

[23] N.Christianini and J. Shawe-Taylor, "AN Introduction to Support Vector Machines and other Kernel Based Learning Methods," Cambridge University Press, United Kingdom, 1st Edition 2010.

[24] S.Bhatti, "Multi Class Sentiment Analysis on Movie Reviews," University of Illinois, Unpublished

[25] T. Fawcett, "An Introduction to ROC Analysis," Pattern recognition Letters, vol. 27, pp. 861-874, 006.

[26] C. M. Bishop, "Pattern Recognition and machine learning," Information Science and Statistics.

[27] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," Information processing and Management, vol. 45, no. 4, pp. 427-437, July 2009