International Journal of Database Theory and Application
Vol.10, No.6 (2017), pp.87-102
http://dx.doi.org/10.14257/ijdta.2017.10.6.07

Capital Markets Prediction: Multi-Faceted Sentiment Analysis using Supervised Machine Learning

Kushatha Kelebeng[1] and Hlomani Hlomani[2]

[1]*Botswana International University of Science and Technology*
[2]*Botswana International University of Science and Technology*
[1]*Kushatha.kelebeng@studentmail.biust.ac.bw*, [2]*hlomanihb@biust.ac.bw*

Abstract

Over the years the stock market has proved to be very difficult to predict due to its unpredictable activities. Data mining techniques such as clustering, decision trees, genetic algorithms and artificial neural networks have been used in order to predict the stock market. Although there has been a significant amount of research done in this area, there are still many issues that have not been explored yet. The impact of fundamental analysis in the prediction of the stock market has been ignored though it can play a vital role in the prediction of the stock market. In this research, the problem of how a social data sentiment correlates to stock price is studied. A stock price prediction model was built using social data sentiments to predict the stock market. Sentiments analysis principles were applied to machine learning techniques in order to find the correlation between the stock market and public sentiments. This study particularly intended to assess the predictability of prices on the Botswana Stock Exchange through the application of Facebook sentiments classification. Three classification models were created that depicted news polarity as happy, calm, alert and vital. Results show that Naïve Bayes and Support vector machine performed well in both types of testing as compared to Random Forest. Naïve Bayes gave good results in terms of error margins with an accuracy of 83.3% making it the best classifier for our data set. When plotting the time series plot of sentiment scores and comparing it to the actual stock price graph, a conclusion can be reached that sentiments and stock prices are related and thus stock prices can be predicted using sentiments.

Keywords: *Sentiment Analysis, Machine Learning, Random Forest, Support Vector Machine, Naïve Bayes*

1. Introduction

Prediction of the stock market is of vital importance to the economy of the country. The Stock market can be affected by a number of factors such as the behavior of investors. Investors can predict how the market will perform using publicly available information such as news. News articles trending on social media such as Facebook about companies listed on a stock exchange can play a very big role when it comes to the movement of a stock because investors do react to news. Previous research on predicting the stock market was based on the Efficient Market Hypothesis (EMH) [1] and the Random Walk Theory of the stock market [2]. According to the Efficient Market Hypothesis, the stock market is driven by new information such as news, other than the past and the present stock prices.

Since social media was introduced, companies in Botswana are also adopting its use (*e.g.*, by using Facebook pages) to reach out to their customers. They use it for marketing and for public relations. As a result of customers interacting with companies on social media, user generated content (UGC) can benefit the businesses. Through the use of data mining tools, opinions of customers can be

ISSN: 2005-4270 IJDTA

retrieved and analyzed. A lot of content on the web indicate the views and sentiments of customers who use the Internet with varying expertise. Investor sentiments coupled with data mining tools can be used to generate estimates of future performances on a lot of issues such as the stock market and even elections. These views are seen on comments and sentiments on social media posts. In this paper we study the prediction power of the sentiments from Facebook on the stock market.

With advances in technology such as social media that encourage sharing and collaboration among the users of the Internet, investors can obtain valuable information about companies that they want to invest in. This information tends to influence the decision of the investors. Recently many researchers have been concerned with the impact of news on the prediction of the stock market [3]. One of the reports found that variance in the stock market could be explained using news. Sentiments of investors can be analyzed and used to predict the stock market [5]. The social mood of investors to a company may be one of the factors that can affect the stock price of that respective company. With the emergence of social media, there is a lot of mood data available. Incorporating historical prices and information from social media can improve the predictive ability of the models [6].

2. Problem Definition and Significance of the Study

The problem this paper addresses is the issue of if and how social media sentiments coupled with supervised machine learning can help with the prediction of capital markets. Various methods have been deployed when it comes to the prediction of capital markets but only a few looked at the issue of the impact of fundamental analysis such as the news and social data on predicting capital markets. The purpose of this study is to find out if social media analysis can be used to predict the company's stock price. The problem to be investigated therefore being can social media analysis be used solely to predict the company's stock price? It is expected that social media has a strong impact on a company's stock price.

According to different theories in the economic segment, when there are disruptions in the segment, economic growth and development will be challenging. Developing countries like Botswana have a growing financial sector which includes the stock market. The financial sector is of supreme importance to the growth of the country's economy, the stock market has an irrefutable impact on the economy and investors benefit a lot from its advantages. In order to be able to contribute to the decision-making process of investors, it is necessary to predict the stock market and as a result mitigate investment risks [7].

Most of the major research on sentiment analysis has been done to predict the polarity of text: positive or negative sentiment, but not subjective opinions along a multi-class continuum. This paper will be focusing on multi-faceted sentiment analysis for stock market prediction.

Moreover, prediction of the stock market has grown to be of more importance as it is determined by the behaviour of investors. The investors also determine stock prices by using information that is available to the public in order to predict how the market will be. News available on social media plays a very big role in influencing the stock price movements as investors react to this news. Research suggests that there is a relationship between social data and the movement of the stock price.

3. Related Work

This section gives a background and a synopsis of previous researches that have been carried out in the areas of social media, investing in the stock market and machine learning. The area of sentiment analysis will also be looked at including

machine learning tools that have been used in the area of sentiment analysis. 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.

3.1. Stock Market and Social Media

According to Fama [8] the efficient market hypothesis states that the prices in an efficient market reflect fully all the available information. If this hypothesis is true, then there won't be any need to predict the stock market using investor sentiments on social media since the information from sentiments will already be reflected on the stock market. Past research has since shown that the stock market is not perfectly efficient [9]. Correlation between social media and stock prices have been studied and researchers used other means of determining investor sentiments even long before social media was used. Furthermore in 2004 Brown and Cliff [10] also looked on the issue of correlation between consumer sentiments and the stock prices. The authors also used consumer data from surveys. The authors then made a conclusion that sentiments had a little correlation with stock prices. The authors used monthly time series and weekly time series for their research. Moreover, Tetlock [11] discussed how sentiments expressed in traditional media influences the stock market. In his findings the researcher states that high media gloom predicts downward pressure on the stock prices and low gloom predicts low market trading volume.

As social media became popular after Facebook was introduced in 2004 then twitter in 2006, studies began to be conducted from these platforms. Bollen et al. in 2011 [12] used data from Twitter to examine how mood states that are extracted from Twitter feeds are correlated to the Dow Jones Index Average (DJIA). The authors stated that the accuracy of predictions could be improved by certain mood dimensions. Their work inspired this research as this paper used the Google Profile of Moods State to improve the mood dimensions.

In 2015 Hu and Dickson [13] implemented their own sentiment analysis algorithm using the random forest and some manually labelled tweets. They used them to investigate the correlation between sentiments of tweets mentioning companies in the DJIA and the stock prices of those companies. The authors found positive correlations for some companies though there was a negative correlation for others. Subha and Nambi [14] also investigated the predictability of the direction of the stock market using past stock market movement data and they concluded that the K-Nearest neighbour performed well in their classification. For the purposes of this paper, we adopt their approach of using machine learning and using social media data to predict the direction of the stock market price changes. Given this background, it shows that improving the prediction of stock prices using social media data is a promising field and there is a lot of research that needs to be done focusing more on using investor sentiments hence this paper.

3.2. Random Walk Theory on Botswana Market

Eugene F Fama defined the random walk theory as the theory that states that stock prices are independent of each other though they have the same distribution [8]. This implies that the past stock trend or movement cannot be used to predict the future movement. Little research has been carried out on the predictability of the Botswana stock market and to test the null hypothesis of the random walk theory. The need for research in emerging and less developed markets like the Botswana market is still not recognized. However, Radikoko [15] observed the effectiveness of the Botswana's stock market by testing whether the random walk theory does

exist in both the domestic and foreign companies' index. The paper studied whether the weak form was valid on the Botswana markets. Various statistical methods were used to assess the efficiency of the market.

The author's results showed that the Botswana stock market is not governed by the random walk theory thus implying that the market is weak-form inefficient. The results meant that investors can easily predict the future stock prices using the historical stock market that they have. The author concluded that using the technical and fundamental analysis on the Botswana stock exchange is not futile. Overall the results suggested that the Botswana stock market returns are predictable.

Furthermore, in 2007 Mollah [16] evaluated the predictability of the Botswana stock exchange daily returns and also tested the null hypothesis of the random walk theory. The daily return series of the Botswana stock exchange for the period of 1989-2005 were used and showed serial autocorrelation of the return series and this meant that the Botswana markets could be predicted. Triangulation econometric approach was employed to assess the predictability of the Botswana stock exchange daily return series. The results nullified the null hypothesis of the random walk theory and clearly indicated the predictability of the Botswana stock market.

Chiwira and Muyambiri in 2012 [17] also carried out a study that evaluated the efficiency of the Botswana Stock Exchange using a number of methods to specifically assess the random walk theory. The random walk hypothesis was rejected; weekly and monthly BSE data was used from 2004 to 2008. It showed that expected stock prices are not independent of past price changes. Though that being the case the market can still be outperformed, both technical and fundamental analysis can yield positive results.

In conclusion, all these studies show that the use of fundamental and technical analysis on predicting the future stock market is worthwhile. This motivates for more research work to be done to continue applying both technical and fundamental strategies.

3.3. Sentiment Analysis

The field of natural language processing is a broad field where the area of sentiment analysis is just a small part. Sentiment Analysis is very popular amongst researchers resulting in a lot of research being carried out in this area [18]. Sentiment analysis is divided into two; the machine learning approach and the lexicon based approach [19]. Both of these approaches have been used to detect sentiments of twitter messages. Nielsen in 2011 [20] used the lexical approach when he investigated the performance of wordlists when doing sentiment analysis for Twitter.

Go *et al.*, [21] showed that Maximum entropy, SVM and Naïve Bayes could have more than 80% accuracy when applied on Twitter for sentiment analysis. This approach has also been used by Pang *et al.*, [18] when comparing the performance of algorithms for the task of sentiment analysis for movie reviews. Pawardhan *et al.*, [22] also followed the machine-learning algorithm approach by using parts of speech and exploring the use of tree kernel to perform sentiment analysis on Twitter Data. Given this background of usage of classifiers for sentiment analysis, we build our algorithms using three classifiers and compared the results in this paper.

3.4. Machine Learning

Machine Learning is a subfield of computer science that evolved from the study of pattern recognition and computational learning theory in artificial intelligence. Machine learning explores the construction and study of algorithms that can learn from and make predictions based on available data. The algorithms deduce a model

from training inputs in order to make predictions on unseen inputs [23]. Machine learning uses artificial intelligence and has a close tie with statistics as both of them use large data sets [24]. Machine learning is also related to the field of pattern recognition. Pattern recognition aims to discover patterns in data. Machine Learning and Pattern recognition are also related to knowledge discovery in databases and data mining. Since the beginning of machine learning a number of algorithms have been built. In this paper classification algorithms will be used. For the classification Naïve Bayes, Random Forest and SVM will be used.

4. Experimental Framework

This section provides an in-depth description of the methodology that was employed in this paper. The research architecture of the paper is mainly in two approaches: a machine learning approach to label the sentiment of each post and an analytical approach to study the prediction power of the sentiment on the stock market. The workflow in Figure 1 was followed to classify social data sentiments for generating stock trend signal.

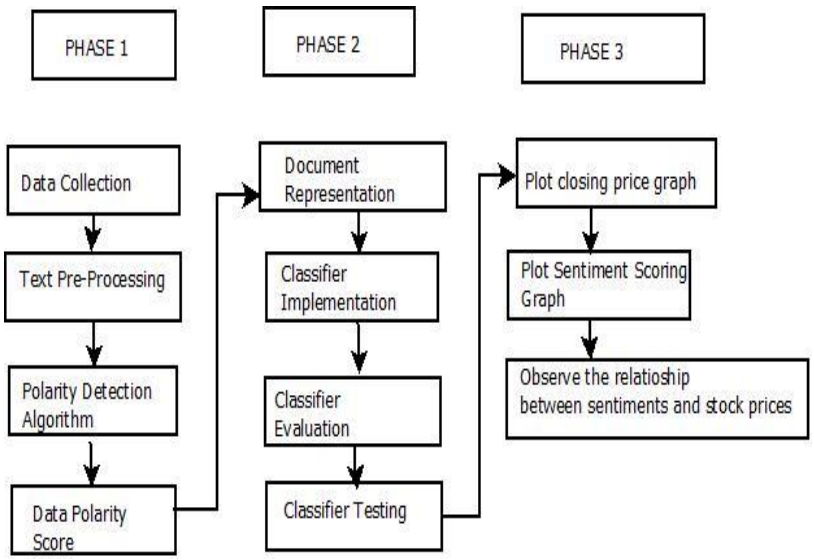

Figure 1. Workflow Depicting the Phases and Steps Taken to Complete the Research

This workflow is in three phases where the result of the first phase is the social data sentiments with their polarity score. These results are passed on to the second phase as input. In the second phase, text is converted (represented) so that it can be feed to the classifiers. Three classifiers are used here so as to compare results. At the end of the second phase results given by all three classifiers are evaluated and tested for classifier performance. In the last phase, the correlation between social data sentiments and the stock market is checked. The data is plotted and the results are recorded. The details of the three phases are discussed below.

4.1. Data Collection

For the purpose of this study a number of things were done, beginning with the collection of data. Two types of data were needed for this project; the time stamped posts and comments from Facebook and stock prices. The following sections show how the data was collected.

4.1.1. Collection of Facebook Data

Facebook being the major social media site was used as the source of data in the experiment. Data was collected from the Official Facebook page of the Choppies Company. This data consisted of publicly available investor posts and comments. The researcher chose Facebook as it covers a wide range of human emotions and captures most of the emotions relevant to sentiment classification. The researcher used 60% of the data collected to train the supervised learning model and the other 40% was used for testing the classifiers. The data was labeled with dates in order to separate it. The company related to the data was also indicated so that relationship between the data and the company‘s stock could be mapped easily. The data collected from Facebook was labeled using only four classes from the Google Profile of Moods State (GPOMS).

4.1.2. Collection of Stock Quotes

There are two Categories in the BSE being the domestic and the foreign companies with a total of 37 stocks or companies listed. The research was focused on domestic companies. The data collected contained the price list of all the 37 companies listed under the Botswana stock exchange but for this research only the Choppies company data was used. The information includes the date, open price, high and low prices and also the trading volume for that day. Data was collected for the period between January 2012 and December 2015.

4.2. Data Pre-Processing

Data from the Botswana Stock exchange was missing for weekends and holidays when the market was closed. In order to make the data complete, missing values were approximated using a concave function. The missing data was approximated by estimating the first day after x to be $(y+x)/2$ to fill the gaps, where x was the BSE value on a given day, and y was the next data point. The reason for using a concave function was that the stock normally follows a concave function unless there is a sudden rise and fall. In addition, for pre-processing the data from Facebook for multi-faceted sentiment analysis the following steps were executed:

4.2.1. Short Word Removal

Short word removal is a process where by all short documents are removed. The minimum length for the document to be removed is given as an input parameter to the function. Documents with irrelevant information or empty documents have a lot of noise. This can be a problem during classification.

4.2.2. Conversion to Lower Case

All words were converted into lower cases in order to remove inconsistency on the use of upper and lower cases. This task does not in any way affect the meaning of words but if it were not done some words would not be considered the same, which could affect the results. It is also important to remove numbers, symbols and white spaces.

4.2.3. Feature Selection

Feature selection is the process of removing redundant features (features, in this case, are words) while retaining those features that have high disambiguation capabilities. Often, some of the features are redundant and noisy. Including them would affect both efficiency and accuracy. For the multi-faceted sentiment classification, words play a major role in determining the sentiment of a sentence. With fewer features, the training

time is reduced and there is less probability for the model to make decisions based on noise.

4.2.4. Tokenization

Tokenization is breaking sentences into meaningful words and phrases. Tokenization divides the text into tokens by separating it using commas, punctuation and whitespaces. After all these have been removed, only numbers and words were left as tokens. Inbuilt libraries in WEKA were used for tokenization (*i.e.*, the string to word vector).

4.2.5. Stop Words Removal

When working with multi-faceted sentiment analysis after feature selection, there are still more words which do not play any role in sentiment classification and have to be removed. The stop words in the Facebook dataset were not only in English but also in other languages and they had to be removed in order to reduce noise in the data. In the Facebook dataset, collected words like in, the, are, as, at, be, by, can, *etc.*, do not determine the overall sentiment of a post or comment. Removing such words increases the performance of the classifier.

4.2.6. Stemming

Stemming is removing the suffix from a word. Stemmers are algorithms that carry out stemming. Their basic function is to trim the words to their original form. For example, if we have the word “paying”, a stemmer will cut the word to “pay”. There are many stemming algorithms like the Porter’s stemmers and the Null stemmer but for this research, a null stemmer was used. In this research, stemmers were basically used to remove the “ng” form, “ly” form and “ed” forms from the words.

4.2.7. The N-Gram Feature

N-grams are “n” continuing words in a sentence. 1 n-gram is known as a unigram, 2 n-grams are known as a bigram and 3 n-grams are known as trigrams. For example, if we have a comment that reads, “Choppies is failing to pay its employees”, the unigrams, bigrams, and trigrams of this comment would be:

Unigrams: {“choppies”, “is”, “failing”, “to”, “pay”, “its”, “employees”}
The Bigrams would be: {“Choppies is”, “is failing”, “failing to”, “to pay”, “pay its”, “its employees”}
While the Trigrams would be: {“choppies is failing”, “is failing to”, “failing to pay”}

N-grams are very important for predicting sentiments. For example, following the example given earlier, “failing” would be predicted as a vital sentiment. N-grams are very useful as they improve the accuracy of the classifier. For the Facebook dataset, unigrams worked effectively as compared to bigrams and trigrams. This is because most investors expressed themselves in a single word like, “bad” and “good”.

4.3. Polarity Detection Algorithm

A post or comment is given polarity based on the amount of positive words versus negative words. Polarity is basically the difference between positive and negative words. If the difference is positive it means the post or comment is positive and if it is negative it means the post or comment is negative. Before the sentiment detection of Facebook posts and comments, Facebook posts and comments were hand labeled as the training and testing set. They were labeled as happy for positive (1), alert for negative (-1), calm for neutral (0), and vital for positive (1). The labeled data are then used to generate training

and testing data set. For some comments and posts it was difficult to label them because they were vague, they were then labeled as calm for neutral.

The following rules were followed in labeling the posts:

- If the post or comment had external links of long articles they were marked as neutral
- Labels were only given when sentiments can be clearly seen from the post or comment.
- Posts and comments with question marks were also marked as neutral.
- Have we seen this post or comment before? If yes, then skip it or use the way it was classified previously.

The results from the manually classifieds data are in CSV (comma separated variable) format with two fields, sentiment: positive, negative or neutral, represented by 1, 0 or -1 and the post or comment.

For automatic sentiment detection of Facebook posts and comments, the dictionary-based approach was followed. The approach uses bag of words technique for text mining. This method is based on the research by J. Bean [25], where he implemented twitter sentiment analysis for airline companies. In these dictionaries, there are list of happy, vital, calm and alert words. The dictionaries are self-compiled. The purpose of the dictionaries is to separate group of words and further to give a way to quantify happy, calm, alert and vital Facebook posts and comments. The compiled dictionaries are based on manually labeled tweets. Four dictionaries were compiled for each class from the Google Profile of Mood state.

Compiling the dictionaries was done following the following simple steps;

1) Import manually labeled posts and comments
2) Take happy posts/comments and extract words from them. This makes the happy words dictionary
3) Repeat step two for calm, alert and vital posts or comments.

Remove duplicate words in all dictionaries. Dictionaries are biased as the datasets are labeled manually, and thus they are based on the personality of the person labeling the dataset.

4.3.1. Polarity Scoring

For scoring sentiments the dictionary-based approach was used. Four types of words collection was used; vital, alert, happy and kind. Each post or comment was matched against these word lists and the number of times it appears in all dictionaries was counted. An assumption was made that both the alert and vital classes make the negative class and the happy and kind classes make the positive class. The posts and comments were then tokenized into a vector of words and a dictionary containing both polarities was created (negative and positive). Each word was then checked to see if it matches either of the dictionaries. The number of words belonging to either positive or negative class was counted. To calculate the score, positive matches were then subtracted from negative matches. If the score is 0 or more the sentiment is considered to be positive (happy or kind), if it is less than 0 then the sentiment is considered negative (vital or alert).

4.4. Document Representation

The documents are represented using the vector space model. Each document is represented using numeric values where each value shows the importance of the document. The TD-IDF method was used in this research. When the vector for document

representation is created, features that are not in the document are assigned the value of zero. The TD-IDF method is very popular amongst researchers since it gives very good results and it is very efficient.

The vector models of text representations were used to train the classifiers. In these models, posts/ comments are represented by words existing in the post/ comment and frequencies. These values are calculated in binary frequency (0 or 1). The Weka platform was used in order to build vector models of unigram (Bag-of-Words), bigram and trigram features for the BSE dataset. Unigrams and bigrams were used to represent posts and comments of the BSE dataset.

4.5. Classifier Implementation

The classifiers that were used for the learning algorithm are the Naïve Bayes, Random Forest and Support vector machines. The algorithms were used to find the sentiments on Facebook sentiments data. They were particularly used to find the relationship between the stock market and the investor sentiments. After all these steps, the data is then divided into two sets; the training and the test set. 60% of the data was used in the training set and remaining 40% was used on the test set.

The implementation stage consists of different processes like data processing, building the model, model operation, evaluating the model and model optimization. The Facebook dataset was first pre-processed using some natural language processing techniques. Three classifiers were used to train the model for comparisons.

The n-fold cross validation procedure was applied for this experiment. In the n-fold cross-validation, the classifier is trained on $n-1$ folds of the data and tested on the remaining fold, then this is repeated n times for different splits, and the results are averaged over the n experiments. For the Facebook dataset, results for 16 cross validation are being reported.

The existing Weka implementation of Naïve Bayes, SVM and Random Forest was used together with their default settings. The Library for Support Vector Machines (LibSVM) for training SVM classifiers with linear kernel and the cost parameter, C, set to 1 was used. Optimizing classifier parameters and using alternative kernels would most likely improve performance and as such can be explored in future studies.

4.6. Evaluation Criteria for System Evaluation

The evaluation criteria chosen play a major role in choosing classifiers. Different metrics are used for the evaluation of classifier performance. In this experiment, accuracy, precision, recall, ROC Area, kappa statistic and mean absolute error were used to evaluate the performance of the classifiers. The precision and, recall metrics evaluate the quality of the algorithms separately for each class (e. g. positive or negative). The Accuracy metric is convenient for multiclass classification tasks to account for imbalanced test data. Since our datasets include Facebook posts/ comments of various polarities, for comparing the performance of different methods the accuracy metric was chosen.

4.7. Methods of Testing

The models were tested using different testing options (10-fold cross validation and the 60% Percentage Split) so that each method can be compared against different cases. It is important to note that model selection is choosing the best model for the task. Cross validation is the most used model selection strategy. To be able to compare the models, many metrics are available to choose from. The following section explains cross validation as it was used in this paper.

4.7.1. Cross-Validation

Cross validation is testing different models over multiple folds on the training set. With cross validation the dataset is divided into equal parts. When training, *k-1* subsets are used and the remaining ones are used for testing. The process is repeated *k* times so that each division is used once as a testing set. The model is trained on part of the training data on each fold. A validation set is then used to test the prediction of the trained model. When all folds have been done the results are combined and compared. Once a training set has been used to build the model, its performance should be evaluated using unseen data; this is called a test set. Since the model is biased towards the training set, a test set is used to avoid high performance measures that are artificial. All the Posts and comments were divided on different sets. For each post and comment, the numbers of words were considered as class labels.

4.8. The Analytical Component

As mentioned earlier, this paper is divided into two parts (a machine learning approach to label the sentiment of each post and an analytical approach to study the prediction power of the sentiment on the stock market). This section explains the last part: the analytical component.

After classification of data, we plotted the sentiment analysis graph and the historical stock price graph. The results from the manually classified data are in CSV (comma separated variable) format with two fields: the sentiment: positive, negative or neutral (represented by 1, 0 or -1) and the post or comment. For comparing the graphs we used descriptive statistics to describe the features of the data.

5. Results and Evaluation

In this section, the results from the experiment are presented and analyzed. As mentioned in Section 4, the research architecture of the paper mainly consists of two approaches: a machine learning approach to label the sentiment of each post or comments and an analytical approach to study the prediction power of the sentiment on the stock market and thus the results are also in two parts.

5.1. Sentiment Detection Algorithm Results

In this study, three classifiers were used, namely: Naïve Bayes, Random Forest and Support Vector Machines. These algorithms were chosen because most recent research shows that these algorithms perform best in text classification. Moreover, the three classifiers were chosen because according to literature the Random Forest classifier has high accuracy and also has good performance. It is also said to be very simple to understand. It is considered best for sentiment analysis [13]. The Naïve Bayes on the other hand were used because the data was not complex and that it is a high bias/low variance classifier. It is good when working with limited amount of data to train a model. Naïve Bayes are simple and can be trained quickly. Support vector machines were chosen because they are very fast to build [13]. All the three algorithms were used to perform multi-faceted sentiment analysis to predict whether a Facebook status update or comment is happy, calm, alert or vital. Around 3000 status updates and comments were collected.

Table 1. Results for the Accuracy Metric

	10 Cross Validation	60% Data Split
Random Forest	66.6%	20.0%
Naïve Bayes	83.3%	80.0%
Support Vector Machines	83.3%	80.0%

Table 1 shows the accuracy results obtained for both the 10 cross validation and 60% data split testing methods. It can be observed that both the SVM model and the Naïve Bayes model are identical in terms of their performance. Both models are superior as compared to the Random Forest on all test methods.

Table 2. Results for the Precision Metric

	10 Cross Validation	60% Data Split
Random Forest	0.708	0.100
Naïve Bayes	0.900	0.867
Support Vector Machines	0.900	0.867

From Table 2, it is comprehensible that the Naïve Bayes and SVM classifiers perform better than the Random Forest classifier in terms of precision in both test methods.

Table 3. Results for the Recall Metric

	10 Cross Validation	60% Data Split
Random Forest	0.667	0.200
Naïve Bayes	0.833	0.800
Support Vector Machines	0.833	0.800

The SVM and Naïve Bayes still outperformed Random Forest classifier again in terms of Recall as depicted in Table 3. Table 3 shows a recall of 0.83 and 0.80 for both SVM and Naïve Bayes classifiers.

Table 4. Results for the ROC Metric

	10 Cross Validation	60% Data Split
Random Forest	0.907	0.867
Naïve Bayes	1.00	0.867
Support Vector Machines	0.935	0.867

Table 4 shows the ROC area results with respect to test dataset. From the table we can see that Naïve Bayes have an ROC Area of 1, which means a perfect test. A perfect test means that the prediction model is good. Both SVM and Random Forest classifiers give almost 0.900 which is closer to 1, that means the models are perfect as well.

Table 5. Results for the Kappa Metric

	10 Cross Validation	60% Data Split
Random Forest	0.556	0.1304
Naïve Bayes	0.778	0.6875
Support Vector Machines	0.778	0.6875

According to Table 5, the highest Kappa Statistics is for both Naïve Bayes and SVM classifier at 0.778, a value that is greater than 0. This means that the SVM and Naïve Bayes classifiers are doing better than chance.

Table 6. Results for the MAE Metric

	10 Cross Validation	60% Data Split
Random Forest	0.2267	0.294
Naïve Bayes	0.0813	0.1501
Support Vector Machines	0.2639	0.2667

Table 6 shows that the NB classifier has the lowest error rate.

According to our experiment results, NB outperforms both the Random Forest and the SVM classifiers in terms of error rates. We can attribute this to the fact that the NB classifier assumes that class attributes within the same class are conditionally independent given the class label.

For Support Vector Machine, the SMO implementation in WEKA was used. For each experiment, the ten-fold cross validation method and the 60% percentage split method was used to evaluate the performance of the SVM classifiers. The results show that SVM perform well as compared to Random Forest. Its best accuracy of 83.3% is achieved by using ten-fold cross validation testing option. SVM does perform extremely well in the ten-fold cross validation as compared to 60% percentage split where it scores only 80%.

For Random Forest, the tree implementation in WEKA was used. For each experiment, ten-fold cross validation method and 60% percentage split method was used to evaluate the performance of the Random Forest classifiers. The results show that Random Forest does not perform well as compared to Naïve Bayes and SVM. Its best accuracy of 66.6% is achieved by using ten-fold cross validation testing option. Random Forest does perform extremely well in ten-fold cross validation as compared to 60% percentage split where it scores only 20%.

With the Naïve Bayes classifier, the Bayes implementation of WEKA was used. For each experiment, ten-fold cross validation method and 60% percentage split method was used to evaluate the performance of the Naïve Bayes classifiers. The results show that Naïve Bayes perform well as compared to Random Forest. Its best accuracy of 83.3% is achieved by using ten-fold cross validation testing option. Naïve Bayes does perform extremely well in ten-fold cross validation as compared to 60% percentage split where it scores only 80%.

As summarized in Tables 1, 2, 3, 4 and 5 the Naïve Bayes and the SVM classifiers give more accurate results for all measures than the Random Forest classifier. For each class the assessment is measured by true positives which are the number of correct examples that are identifies as part of the class. It is also measured by true negatives, which are correct examples that we not recognised as part of the class. They are also measured by false positives (examples that were incorrectly assigned to the class) and false negatives (examples that were not recognised at all).

In existing literature, there is no downward trend for the performance of a Random Forest Algorithm in any available data sets. Studies show that the performance of the Random Forest is always rising. They are known to be one of the most efficient classifiers. In our experiment, the Random Forest has a slightly lower performance than both the Naïve Bayes and the SVM. This is contrary to existing literature. The reason why the Random Forest performed worse might be because of the number of features we had. The Random Forest classifiers perform well when there are many features. The other reason may be due to the complexity of data distribution. A random forest may give wrong classification results due to the inclusion of bad tree classifiers. This bad performance may also be attributed to the situation where the correlation between trees was not minimised. In some case, the occurrence of correlated trees may also affect the performance of the Random Forest resulting in it not getting good results.

Comparing the Naïve Bayes to Random Forest we can see that Naïve Bayes produces a high performance. This is because the Naïve Bayes is a probabilistic classifier that is based on the Bayes theorem with strong independence assumptions. Being one of the most basic text classifiers it is convenient because it can be trained fast and it also performs well in complex tasks such as sentiment analysis.

5.2. Facebook Data Sentiments vs. Stock Closing Price

Sentiment data was formatted so that it fits to the existing format of the stock data. The sentiment data was then used to create a trend form. To find out if indeed there is a correlation between stock prices and social media data sentiments, the sentiment polarity graph and the stock market graph are compared by looking for similarities and drawing conclusions. Drawing conclusions was the most difficult part. The comparison informed conclusions on the trustworthiness of the sentiment trend. A conclusion was made that traders can trust in sentiment analysis as a new way of predicting trends and as a trend indicator of the stock market.

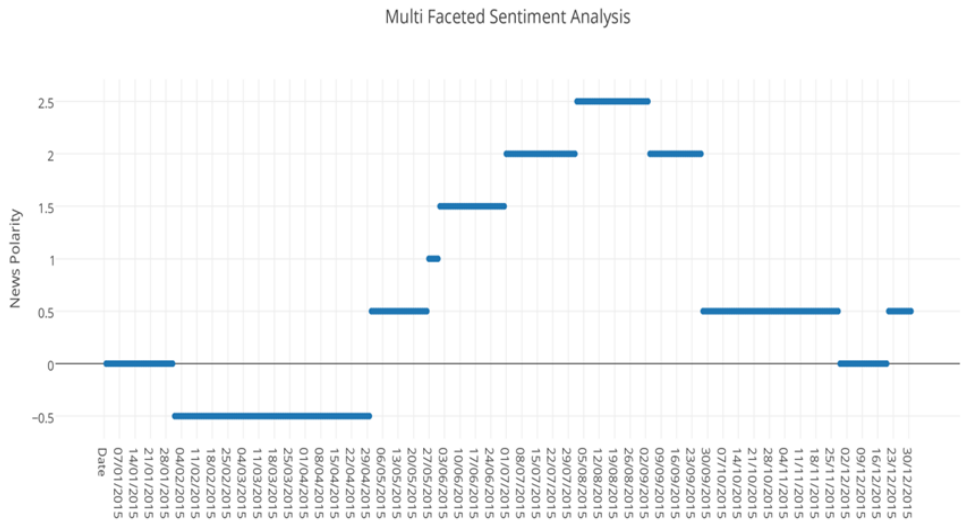

Figure 2. The Graph for Multi-Faceted Sentiment Analysis

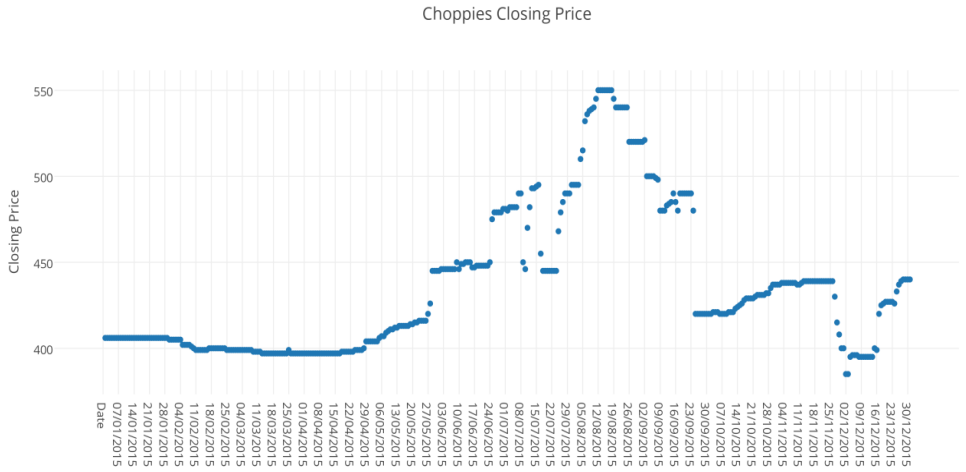

Figure 3. The Graph for Choppies Closing Stock Prices

In both Figures 2 and 3, a similar pattern is occurring between the two graphs. The overall direction of the two graphs is similar. There seem to be a non-linear trend over both the sentiment polarity and the closing prices. There are points that have to be taken into consideration for the analysis of these plots. There are several areas on the stock market graph that are flat, as they show no variance and show a sudden rise or fall afterwards. This is because of the way the stock market works; the stock market is only open during the week and closed on holidays and weekends. During the time when it‘s closed they won‘t be any data on closing prices. So for this study a concave function was used to fill in the missing values. This accounts for the flat areas of the stock price plot. As for the sudden rise and fall of the stock market, it is due to the fact that even when the stock market is closed, people can still buy and sell and a computer program then computes an opening price for the next day, which can be different from the actual closing price of the last day.

On the sentiment analysis graph, most of the areas are flat and show no variance and then show a sudden rise and fall afterwards. This is because the Botswana stock market is a developing market and its investors do not spend much time on social media discussing companies listed on the market.

6. Conclusion

Over the years, researchers from different domains have studied the stock market. Online social media emerged over the years as a novel form of communication. We believe that its data can be used to carry out analysis. Through social media, people obtain information about a decision by communicating with one another and social media has become an important aspect of decision-making. This paper used a unique approach of Facebook posts and comments multi-faceted sentiment analysis to find the correlation between stock market and social media sentiments. Sentiments are quickly reflected in the stock market whether they are positive or negative. The lexicon-based approach has been used by most previous work in this area. In this paper, sentiment analysis was done using machine learning algorithms in natural language processing in order to study the correlation between stock prices and social data sentiments.

Stock trends depend on a number of factors and predicting its future trends is a vital task. Social media sentiments and stock prices are related to each other and sentiments from investors on social media have the capacity to fluctuate the stock trend. The relationship or correlation between the two were studied and a conclusion was reached that stock trend can be predicted using social media data sentiments and past stock price

history. Social media data show how the investors feel about the current market; the sentiment detection algorithm was automated based on the words on the posts and comments to get overall polarity. If the sentiment is positive then there is more chance of the stock price going high, as the sentiment impact is good on the market. If the sentiment is negative then it will impact the stock price to go down.

7. Recommendation

Over the years researchers from different domains have studied the stock market. Online social media emerged. There were a number of limitations in this research. Data was collected only for a period of twelve months; the results could have been more significant if the data could have been collected over a longer period. Moreover, including many companies that are listed on the BSE can be done in further research. For this research, only one company was used, as it is the only one active on social media at the moment. Lemmatization can also be used as a possibility to improve this work. Lemmatization is the same as stemming but it replaces the suffix of a word with a typical word suffix to get the normalized word form rather than generating a stem of the word. In this study financial words were not included in the dictionary but adding financial words could lead to even better results. Moreover, the prediction performance can be improved by improving the way data is presented and prepared when it is used as input to the prediction models. This can be achieved by using Trend Deterministic Data Preparation Layer and Two Stage Fusion Models for the task of predicting the direction of movement and value. A single dataset is not enough for testing and comparing different machine learning algorithms. The algorithms should be tested using different datasets so as to measure the performance of the algorithms.

Acknowledgments

I would like to thank Dr Hlomani Hlomani for having been a great mentor during the entire process of writing this paper.

References

[1] M. A. Bayir, H.I. Toroslo and G. Fidan, “Smart Miner: A new Framework for Mining Large Scale Web Usage Data”, **(2009)**, pp. 161-170.
[2] C.T. Li, C.Y. Wang, C.L. Tseng and D.L. Shou, “A Sentiment-based Audiovisual System and Displaying Micro Blog Messages”, 49th Annual meeting of the Association for Computational Linguistics: Human Language Technologies, System demonstrations, Portland, Oregon, **(2011)**, pp. 32-37.
[3] F. Li, “Do Stock Market Investors Understand the Risk Sentiment of Corporate Annual Reports”, working paper, **(2006)**.
[4] M.A Mittemayer and G. Knolmayer, “NewsCats: A News Categorization and Trading System”, 6th International Conference on Data Mining, **(2006)**, pp. 1002-1007.
[5] T. Nguyen, “Topic Modeling based Sentiment Analysis on Social Media for Stock Market Prediction”, **(2015)**, pp. 1354-1364.
[6] Y. Kara, M.A. Boyacioglu and O.K. Baykan, “Predicting direction of stock price index movement using artificial neural and support vector machines: The sample of Istanbul Stock Exchange”, Expert Systems with Applications, vol. 38, **(2011)**, pp. 5311–5319.
[7] E.F. Fama and M.E. Blume, “Filter rules and stock market trading”, Journal of Business, vol. 39, no. 1, **(1966)**, pp. 226-241.
[8] J. Grossman and E. Stiglitz, “On the impossibilities of informationally efficient markets”, The American economic review, vol. 10, no. 3, **(1980)**, pp. 393-408.
[9] G. W. Brown and M. T. Cliff, “Investor Sentiments and the near team stock market”, Journal of Empirical Finance, vol. 11, no. 1, **(2004),** pp. 1-27.
[10] P. C. Tetlock, "The role of media in the stock market," Journal of Finance, vol. LXII, no.3, **(2007)**, pp. 1139-1168.
[11] J. Bollen, H. Mao and X. Zeng, “Twitter Mood Predicts the Stock Market”, Journal of Computational Science, vol. 2, no. 1, **(2011)**, pp. 1-8.

[12] B. Dickson and W. Hu, "Sentiment Analysis on Investor Opinions on Twitter",Social Networking, vol. 4, **(2015)**, pp. 62-71.
[13] M. V. Subha and S. T. Nambi, "Classification of stock index movement using K-Nearest Neighbour", Information Science and Applications, vol. 9, no. 9, **(2012)**.
[14] I. Radikoko, "Testing the Random Walk behavior of Botswana's Equity Returns", Journal of Business Theory and Practice, **(2014)**, pp. 84.
[15] 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, **(2007)**, pp. 1077-1094.
[16] 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, **(2012)**, pp. 83-91.
[17] B. Pang, L. Lee and S. Vaithyanatha, "Thumbs up? Sentiment classification using machine learning", Proceedings of the Conference on Empirical methods in natural language processing, **(2002)**, pp.79-86.
[18] P. Goncalves, "Comparing and Combining Sentiment Analysis Methods", COSN 2013, **(2013)**
[19] S. Nielsen, "state of the media: the social media report", Q3 **(2011)**.
[20] A. Go, R. Bhayani, and L. Huang, "Twitter Sentiment Classification using distant Supervision", Stanford, CS229 Project Report **(2009)**.
[21] S. Patwardhan, B. Bugaraev and A. Agarwal, "Labeling Landscaping: Classing tokens in context by pruning and decorating Trees", CIKM'12, **(2012)**.
[22] P. Kohavi and F. Provost, "Applications of Data mining to Electronic Commerce, Discovery Process", vol. 30, no. 2/3, **(1998)**.
[23] L. Breiman, "Statistical Modeling: Two Cultures", vol. 16, no. 3, **(2001)**, pp. 199-231.
[24] J. R. Bean, "By example: mining Twitter for consumer attitudes towards airlines, Boston Predictive Analytics meet up presentation", **(2011).**

Authors

Kushatha Kelebeng, Kushatha Kelebeng is an MSc Computer Science student at Botswana International University of Science and Technology, and also a teaching assistant in the Department of Computer Science and the Department of Technical Writing and Academic Literacy. Kushatha holds a Bachelor's Degree in Computer Systems Engineering from the University of Sunderland. Her research interest includes social media data mining, big data mining & analysis and mobile studies.

Dr Hlomani Hlomani, Hlomani Hlomani received his bachelor's degree in Information Technology from the Cape Peninsula University of Technology, South Africa in 2005 and both his MSc and PhD degrees at the University of Guelph, Canada in 2009 and 2014, respectively. Currently he is a Lecturer and Researcher in the College of ICT at the Botswana International University of Science and Technology. He holds an adjunct position in the school of Computer Science at the University of Guelph. His research interests span across several disciplines within Computer science domain including knowledge engineering, artificial intelligence, distributed systems and software engineering. He basically seeks solutions to computing problems using a blend of different technologies. Recent research problems include those in data integration, modelling, knowledge representation, prediction, clustering, classification and machine learning. He has received the best paper award at the Knowledge Engineering and Ontology Development (KEOD) Conference in 2014. He has published several papers including journal articles, conference papers and book chapters.

