I.J. Information Technology and Computer Science, **2016, 10, 11-21**
Published Online October 2016 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2016.10.02

Prediction Model of the Stock Market Index Using Twitter Sentiment Analysis

Anthony R. Caliñgo
Technological Institute of the Philippines, Quezon City, 1109, Philippines
E-mail: anthonycalingo@me.com

Ariel M. Sison, Bartolome T. Tanguilig III
Technological Institute of the Philippines, Quezon City, 1109, Philippines
E-mail: ariel.sison@eac.edu.ph, bttanguilig_3@yahoo.com

***Abstract*—**Stock market prediction has been an interesting research topic for many years. Finding an efficient and effective means of predicting the stock market found its way in different social networking platforms such as Twitter. Studies have shown that public moods and sentiments can affect one's opinion. This study explored the tweets of the Filipino public and its possible effects on the movement of the closing Index of the Philippine Stock Exchange. Sentiment Analysis was used in processing individual tweets and determining its polarity - either positive or negative. Tweets were given a positive and negative probability scores depending on the features that matched the trained classifier. Granger causality testing identified whether or not the past values of the Twitter time series were useful in predicting the future price of the PSE Index. Two prediction models were created based on the p-values and regression algorithms. The results suggested that the tweets collected using geo location and local news sources proved to be causative of the future values of the Philippine Stock Exchange closing Index.

***Index Terms*—**Social media, sentiment analysis, causality, data mining, stock market.

I. Introduction

Decision-making is a vital part of our daily lives. For us to make wise judgments, we mostly rely on the past events, other people's opinions, or just plain observation. There is an undeniable fact that knowledge and awareness are missing in people investing in the stock market, that's why prediction methods are very important in enticing people to participate in trading, as well as, to retain existing investors. Stock market investors put a lot of money in companies they are not associated with, and mostly, based on instincts and word of mouth [1]. The growth of an economy in a country is relative to the performance of the stock market; It is also said that the stock market is driven by its investor. In this sense, forecasting has been a great interest in the stock market because it can serve as a guide for traders and investors - just like in weather forecasting.

With the fast growing number of social networking platforms, public opinions [2] started to play a bigger role in financial markets, and as of 2008, nearly one in four adults in the US were reported to rely on social media channels for investment advice [3]. Online communities and social media (e.g. Twitter and Facebook) also play a huge role in influencing investments made by people [4]. Twitter and Facebook land in the top 10 most visited websites in the world [5]. Since there is a huge amount of people expressing personal opinions, it is safe to assume that these social media platforms can be one great source of information [6][7]. With proper tools and the help of technology, meaningful and precious information can be gathered, analyzed, and utilized in different areas like in the movement and performance of the stock market.

Prediction in social media analyzes information gathered based on a user's opinions and beliefs [5]. This information is then compared with facts and data in determining if the prediction is accurate or not.

Generally, we usually just follow or copy the activities and actions of others, which is a common mistake in investing. Since the stock market is dictated by its investors, sentiments of the people can be a factor in its day to day performance.

This study, therefore, examined whether social media has a significant predictive relationship with the daily performance of the Philippine Stock Market Index (PSEi). It utilized public tweets from individuals, several local news sources in the Philippines like ABS-CBN News (@ABSCBNNews), GMA News (@gmanews), and relevant data using hash tags and keywords. It evaluated the Twitter data and transformed them into meaningful information using Naïve Bayes algorithm for sentiment analysis and regression algorithms for the Granger causality test, which was used to create a prediction model.

The rest of the paper is organized as follows: Section II presents other works in relation with the study. Section III describes the design concept of Sentiment Analysis and Granger Causality. Section IV details the methods of experimentation and algorithms that were used to produce the prediction model. Section V presents the output of the algorithms and result of the sentiment analysis and prediction models. Finally, the conclusion and future improvements of the paper are discussed in Section VI.

II. Related Works

Mayfield described social media as a group of new kinds of online media, which share characteristics of participation, openness, conversation, community, and connectedness [8]. Kwak et al. [9] conducted a study in order to find out how Twitter is disseminating information. They identified how the topologies (follower-following structure) affected the transfer or information. They also concluded that trending topics usually come from news sources amounting to 85% coming from news headlines. It was also revealed that re-tweeted tweet reaches an average of 1,000 users factoring out the number of followers of the source.

In the work of Asur and Huberman [10], they showed how to forecast future outcomes, specifically the box-office revenues of movies before their release date. They utilized the chatter from almost 3 million tweets from Twitter and used linear regression model for their prediction. The results outperformed the accuracy of the Hollywood Stock Exchange and that there was a strong correlation between a movie's ranking in the future. Tumsajan et al. [11] and Bermingham and Smeaton [12] studied the predictive power of social the Twitter platform in predicting the outcome of the political elections. They found a fairly significant positive results but both studies concluded that Twitter, or social media, alone doesn't have the ability to give a high percentage of predictive power regarding election results.

Bollen et al. [13] explored how public mood patterns relate to fluctuations in macroscopic social and economic indicators in a given period. They performed a sentiment analysis using the Profile of Mood States (POMS) of all tweets published in the second half of 2008 and discovered that events in the social, political, cultural, and economic domain have a significant effect on the various dimensions of the public mood.

Mittal and Goel [14] used Twitter data to predict public mood and used DJIA values to forecast stock market movement. They proposed a new cross-validation method for financial data and obtained 75.56% accuracy using Self-Organizing Fuzzy Neural Networks (SOFNN).

Their works were followed by Zhang, et al. [15] who tried to predict the DJIA, NASDAQ, S&P 500, and VIX by analyzing Twitter posts. Rao and Srivastava [16] who got an 88% accuracy based on their sentiment analysis using Twitter. Ding et al. [17] as well conducted a study but only achieved 51.88% accuracy because of the method they used in collecting their Twitter data.

III. Design Concept

The study went through several steps in order to create a prediction model that investigated the causality of the Twitter data to the stock price movement. This section presents the main concepts and procedures that were followed by the researches from Sentiment Analysis to Granger Causality.

Fig. 1 shows the framework of the study. It is divided into four main parts: 1) Data Extraction 2) Pre-Processing 3) Sentiment Analysis and 4) Granger Causality Analysis and Prediction. Below is a discussion of each step to clearly give an idea of the processes that were be involved in the study.

A. Data Collection

The first part was the collection of data from Twitter and the stock market. These two served as the sources of information, which were used as the training samples for this study. The tweets were collected by using Twitter's API and Python. All scripts for extracting the tweets were written and executed using Python programming language. Python has readily available packages and libraries that can be easily accessed in order to perform several processes that need to make use of APIs over the Internet.

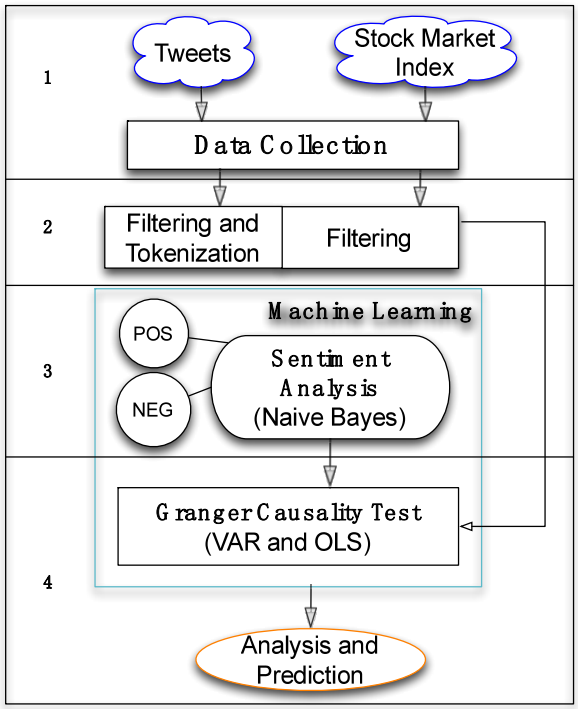

Fig.1. Operational framework of the study.

This research explored public tweets and their sentiments in three different ways:

1. Public tweets in the Philippines by the use of a geocode:
   Latitude: 14.589119422692292
   Longitude: 121.0263763730469
   Radius: 17097.55 KM
2. Tweets containing hash tags and keywords that are specific for the stock market audience:
   -"PSEi", "PSEindex", "Philippine Stock Market", "Philippine Stock Exchange"
3. Tweets from top local news sources (users):
   -"@ABSCBNNews", "@ANCALERTS", "@PhStockExchange","@Gmanews","@Cnnphilippines","@philippinestock","@inquirerdotnet","@PhilstarNews", "@manila_bulletin", "@bworldph", "@BusinessMirror"

The data were divided into three different analyses in order to find out if whether there was enough useful information coming from the general public tweets, hash tags or specific users in predicting the movement of the stock market.

The data gathered for this study were from June 2 to August 31, 2015 (91 days) with an estimated number of 800,000 – 850,000 tweets or 9,300 tweets per day. The stock market closing Index data was downloaded from Quandl.com.

B. Pre-Processing

The collected data now goes to the second step, which is the Pre-Processing. Fig. 2 outlines the pre-processing steps for the Twitter data and stock market data. The pre-processing stage translated all tweets to English using GoSlate API in Python. The API provided access to Google's online translation via a python script. The Python script read all gathered tweets and translated them into English. Since the translation was done using Google's algorithm, grammatical errors cannot be avoided and the context of the tweet might be lost in the process. This is alright because the classifier in sentiment analysis used a bag-of-word approach in extracting features from a given set of words. The pre-processing of the stock Index data only required preparation of the closing values of the PSEi for the time series.

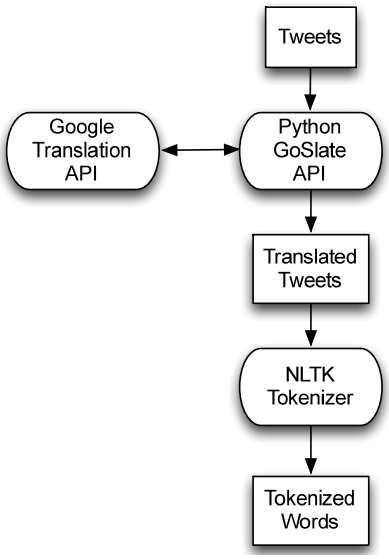

Fig.2. Pre-processing flow of Twitter data.

After translating the tweets to English, they needed to be tokenized first before the actual sentiment analysis can begin. Machines cannot understand the human language and its corresponding emotions, so the tweets need to be converted first into a format in which machine algorithms can be performed. The NLTK tokenizer performed these tasks and broken down the tweets into chunks in order to have a more accurate classification process.

E.g. Tokenization Using NLTK Tokenize Package

String = #PSEi lackluster in morning trade: down 2.29 pts or 0.03% to 7,720.85; regional marts tempered by jitters over Greece

Tokenized = ['#', 'PSEi', 'lackluster', 'in', 'morning', 'trade', ':', 'down', '2.29', 'pts', 'or', '0.03', '%', 'to', '7,720.85', ';', 'regional', 'marts', 'tempered', 'by', 'jitters', 'over', 'Greece']

The tokenization process prepared the text from each tweet into desired individual parts: words, punctuations, letters, and special characters. For this study, the researchers used an open source Python library called the NLTK Tokenizer Tool (nltk.org). Tokenization created a bag-of-words collection that was used for feature extraction in sentiment analysis.

C. Sentiment Analysis

The third step involved the sentiment analysis process, which is a type Natural Language Processing (NLP), with the intention of getting sentiment or subjective information from a given text [18]. NLP is a type of computer manipulation done in a natural language like English and Filipino. Text analysis enables us to detect sentiments in sentences, or specifically, Tweets.

Sentiment analysis in the context of NLP involves the analysis of comments left on social media sites like Twitter. But, instead of analyzing just words, sentiment analysis identifies the person's attitude towards a something by using variables and features. These sentiments can be classified and transformed into meaningful information that can be used for a variety of purposes such as prediction.

In this step, tweets were polarized into positive and negative and were given a score between 0 and 1. Each tweet was given a negative and positive score, which equates to 1 when added, and whichever has the score higher, will dictate its polarity. Positive and negative are called sentiment (opinion) orientations or polarities.

D. Granger Causality

The fourth step is for the prediction modeling and analysis of the daily movement of the stock market Index (PSEi). This step tried to find of there was a causative relationship between the Twitter sentiments and the stock market, or if it only shows mere a correlation. It will use the Granger causality analysis that was introduced by Clive Granger [19]. This is based on the linear regression algorithm [20] to determine the causality of the generated time series from sentiment analysis scores and the closing Index of the stock market. Granger causality doesn't imply true causation but instead, tests if one variable is helpful in predicting another variable. P-values were used to determine if a null hypothesis can be accepted or rejected.

Economists use Granger causality as a tool to investigate a statistical pattern of lagged correlation. In this method, time series X is said to cause time series Y, if it can be proved that time series X provide statistically significant information about the future values of time series Y, than Y alone [21]. The lagged values of time

series X will have a statistically significant correlation with time series Y. It tested whether Twitter sentiments (pos, and neg) has a causative effect ("Granger causal") on the movement of the stock market Index. To select the optimal lag value, this paper used the Akaike Information Criterion (AIC) to measure the quality of a model [22].

IV. Methods and Experimentation

This section discusses the methods and algorithms that were used in sentiment analysis, Granger causality and creating the prediction model.

4.1 Sentiment Analysis Classification

The purpose of Sentiment Analysis as can be seen in Fig. 3, is to automatically classify a tweet as either positive or negative, based on a set of features and in trained classifier. The classifier was trained first using the movie review corpus, which is readily available in the NLTK package in Python. This data set contains 1000 positive and 1000 negative movie reviews, which were used in training the Naïve Bayes algorithm [23][24].

The movie review corpus was used because contains a collection several positive and negative sentiment words that were helpful in training the classifier. Since the Naïve Bayes algorithm uses a bag of word approach in classification, the entire context of one review was disregarded and it only collected specific words that correspond to a high positive or negative sentiment probability. Such collection of words is helpful in training a classifier because the reviews contain actual user emotions and sentiments about a movie which can be used in determining whether the sentiment of a tweet is positive or not.

A list containing word features was generated, with individual words in the frequency distribution. After the features had been extracted, the classifier was trained using the NTLK Naïve Bayes classifier. Equations (1) and (2) define the algorithm used by the classifier.

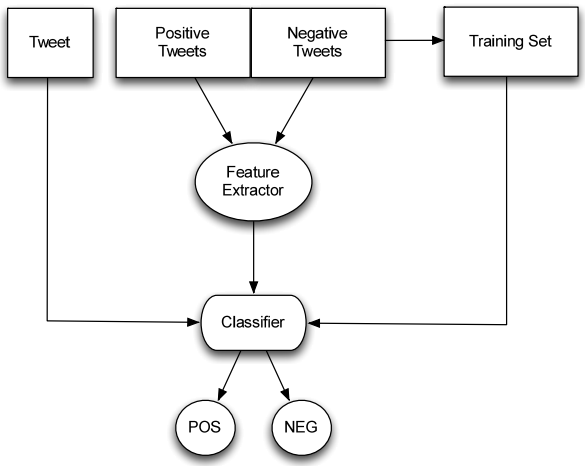

Fig.3. Naïve Bayes Sentiment Analysis.

$$Y_{nb} = argmax\ P(c_i)\Pi P(x_j|c_i) \quad (1)$$

$$p(c_i|x) = \frac{P(c_i)\Pi P(x_j|c_i)}{P(x)} \quad (2)$$

Where:

P(ci | x) is the posterior probability of class c, given predictor x

P(ci) is the prior probability of class

P(x | ci) is the likelihood or the probability of the observation given the class

P(x) is the prior probability of predictor

This means that the most likely class is the class that maximizes the a) the product of the prior probability of the class and b) the product over all the attributes of the product of the attributes given the class. This classifier was used to determine the sentiment of a single tweet and the overall collective polarity of tweets in a certain trading day.

The classification algorithm in the Python NLTK package makes use of this equation:

$$P(lab|feats) = \frac{P(lbl)*P(f1|lab)*...*P(fn|lab)}{SUM[1](P(l)*P(f1|l)*...*P(fn|l)} \quad (3)$$

The NLTK Naïve Bayes classifier in Equation (3) applies prior probability of each word or label and how many times it appeared in the frequency distribution list. This means that if a tweet's polarity needs to be determined, the classifier will look at the training data and multiply each score whenever a word in a tweet appears in the frequency distribution and decide whether it is positive or negative.

Table 1 shows that the first tweet received a score of 0.8771 or 87% negative probability and 0.1229 or 12% positive probability. The second tweet, on the other hand, got a score of 0.4789 or 48% negative probability and 0.5211 or 52% positive probability. Based on the results, the first tweet, therefore, was predicted to be negative and the second tweet to be positive.

This test used random samples from the pre-processed set of Twitter data. Each tweet was analyzed using the NLTK Naïve Bayes classifier, which used the maximum probability of a class, given the set of features, whether it can be classified as positive or negative. The classifier correctly identified the polarities of the test data with their corresponding polarity scores.

Table 1. Test Results of Sentiment Analysis Classification.

Tweet	Negative Probability	Positive Probability	Sentiment Result
I do not really know in the heat of the day!!! The poison?	0.8771	0.1229	Negative
I love this movie	0.4789	0.5211	Positive

All the data gathered from Twitter was analyzed, and the positive and negative polarities were generated. To produce the time series, the positive and negative ratio of the tweets needed to be calculated first. The classification produced a daily summary of the total number of negative and positive tweets. The two summations were used in

the computation of the ratio and these produced the values of the time series.

Equation (4) shows the computation of the polarity ratio. These ratios will then be used to produce the different time series.

$$PNRatio = \frac{Positive\ Tweets}{Negative\ Tweets} \quad (4)$$

Figs. 4, 5 and 6 show the three different time series of the tweets gathered from geo location, hash tags, and users/news sources.

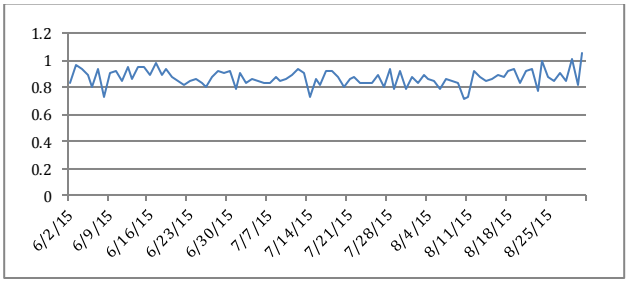

Fig.4. Geo Location Time Series.

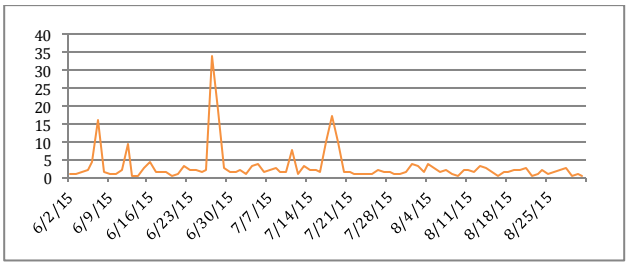

Fig.5. Hash Tag Time Series.

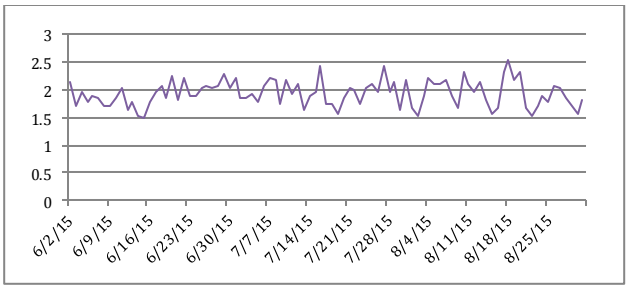

Fig.6. User Time Series.

The time series were based the overall positive and negative polarities of tweets gathered from June 2 to August 31, 2015.

4.2 Granger Causality Test

Granger causality is the main component in achieving a prediction model for this study. It tested whether the past values (lags) of certain time series is useful in predicting the future outcome of another. In order to build relations from multiple time series, different approaches were used to accept of reject the null hypothesis (H0), to all possible pairs of time series. The H_o is:

H_o = Tweets does NOT Granger Cause PSEi

The above null hypothesis can be rejected by looking at the p-values that will be produced by the Granger causality test. The significance level is usually set at 0.05 or 5% confidence level. Therefore in order to say that the tweets Granger cause the PSEi, the null hypothesis must be rejected.

Granger causality has nothing to do with the concept of causality in the philosophical sense. For example, lightning precedes rain. But in reality, lightning does not cause rain. Granger causality is therefore related to the usefulness of one variable in forecasting another variable.

Equation (5) shows a regression of y_t on lagged x_t and lagged y_t , x_t does not cause y_t if all the coefficients on the former are zero. In the regression:

$$Y_{t=} \sum_{i=1}^{p} \alpha_i Y_{t-i} + \sum_{i=1}^{p} \beta_i x_{t-i} + u_t \quad (5)$$

X is said to Granger Cause Y if it satisfies Equation(6):

$$E(Y|Y_{t-k}, X_{t-k}) \neq E(Y|Y_{t-k}) \quad (6)$$

Where Y_t is the current value of Y , Y_{t-k} is the past/lagged values of Y, X_{t-k} is the past/lagged values of X. One time series is said to be Granger causal if the lagged regression of X is helpful in predicting the futures values of Y, than Y alone.

4.2.1 Optimal Lag Selection

Granger causality uses the past values or lags of two time series to perform a regression algorithm. The lagged time series of both the Twitter data and PSEi were needed as dependent and independent variables for the regression model. Not knowing the best number of lags to include may result in overfitting because it might include too many parameters, which might result in a poor predictive performance. Having too few or too many lags may affect the accuracy of the Granger causality analysis. In order to determine the number of lags to be used, the Akaike Information Criterion (AIC) was used in the selection of the optimal model for the Granger causality test.

The AIC, as defined in Equation (7), is calculated as:

$$AIC = n * \ln\left(\frac{RSS}{n}\right) + 2 * K \quad (7)$$

Where n is the number of observations, RSS is the residual sums of squares, K is the number of parameters.

Based on the AIC on Table 2, the optimal lags for the three time series are 2 Lags for geo location and hash tags and 3 Lags for users. These lags obtained the smallest possible value for the AIC.

4.2.2 Pairwise Granger Causality

This study used a Pairwise Granger causality test, which checked for causality in both directions, by using Vector Autoregression Regression (VAR) and Ordinary Least Square (OLS), in estimating the coefficients. F-statistics and p-values were used in telling whether one time series causes the other of if the H_0 can be accepted or rejected.

Table 2. Akaike Information Criterion Results

Akaike Information Criteria					
Sample: 1 91					
Included observations: 83					
Geo location		**Hash tags**		**Users**	
Lag	**AIC**	**Lag**	**AIC**	**Lag**	**AIC**
0	10.69928	0	19.42405	0	13.46876
1	8.019759	1	16.60134	1	10.74092
2	**7.928150***	2	**16.49675***	2	10.65235
3	7.964756	3	16.5391	3	**10.57197***
4	8.056879	4	16.5994	4	10.64661
5	8.077186	5	16.67219	5	10.6742
6	8.151957	6	16.74576	6	10.73801
7	8.225526	7	16.83193	7	10.78977
8	8.293743	8	16.91487	8	10.86559

Null Hypotheses:

1. TRGeo does not cause PSEi
2. PSEi doest not cause TRGeo
3. TRHash does not cause PSEi
4. PSEi does not cause TRHash
5. TRUser does not cause PSEi
6. PSEi does not cause TRUser

A VAR is a multiple variable autoregressive model and the repressors are lags of the same variables in the model. VAR models were used for the analysis of multivariate time series. The structure of VAR is each variable is a linear function of lags (past values) of itself and lags of another variable. This model is useful in showing the behavior of economic and financial time series and can be helpful in forecasting. Since VAR was used to test Granger causality, only two variables were included: Twitter and PSEi time series.

$$Y_t = \alpha_1 + \delta_1 t + \phi_{11} Y_{t-1} + \cdots + \phi_{1p} Y_{t-p} + \beta_{11} X_{t-1} + \cdots + \beta_{1q} X_{t-q} + e_{1t} \quad (8)$$

$$X_t = \alpha_2 + \delta_2 t + \phi_{21} Y_{t-1} + \cdots + \phi_{2p} Y_{t-p} + \beta_{21} X_{t-1} + \cdots + \beta_{2q} X_{t-q} + e_{2t} \quad (9)$$

Equations (8) and (9) show a VAR model with two variables or predictors. Each variable is a linear function of the lag 1 values for all variables in the set. In the X_t model, the lag 2 values for all variables are added to the right sides of the equations, In the case of two variables (or time series), there would be four predictors on the right side of each equation, two lag 1 terms and two lag 2 terms.

Since this is a test for pairwise Granger causality, both variables were tested against each other, with a significance level of 0.05 (p-value). Results are shown in Table 3.

The pairwise Granger causality shows a high level of significance with the geo location and user time series with p-values of 0.0376 and 0.0051. Therefore, we can reject the null hypothesis and say that geo location and user time series Granger cause the PSEi. The hash time series failed to reject the null hypothesis with a p-value of 0.8305.

Table 3. Pairwise Granger Causality Results

Null Hypothesis:	Obs.	F-Stat	Prob.	Remarks
TRGEO does not cause PSEI	89	3.4109	**0.0376***	Reject
PSEI does not cause TRGEO	89	1.99854	0.1419	Accept
TRHASH does not cause PSEI	89	0.18611	0.8305	Accept
PSEI does not cause TRHASH	89	1.40236	0.2517	Accept
TRUSER does not cause PSEI	88	4.59396	**0.0051***	Reject
PSEI does not cause TRUSER	88	0.96347	0.4141	Accept

*p-value <= 0.05 (5%)

The results also show that the Granger causality only goes one way. PSEi does not Granger cause the other variables' time series because they fail to reject the null hypothesis with p-values of 0.1419, 0.2517 and 0.4141 for each time series test.

The f-value was used in deciding whether the model is statistically significant in prediction, whether the regression sum of squares is large enough with the number of variables given. The f-value is the ratio of the mean square of the model and the error mean square. The null hypothesis is that if all population regression coefficients are 0, the model has no predictive capability. The null hypothesis is then rejected if the f-value is large. The general equation for the f-value is defined in Equation (10) as:

$$F = \frac{MSR}{MSE} \quad (10)$$

Where:

MSR = Regression Mean Square or the explained variance

MSE = Mean Square Error or the unexplained variance

If $\beta 1 = 0$, then we'd expect the ratio MSR/MSE to equal 1.

If $\beta 1 \neq 0$, then we'd expect the ratio MSR/MSE to be > than 1

Figs. 7, 8 and 9 show the F distribution of the different time series with their corresponding F-critical values. The critical value is the number that the F-value must exceed to reject the null hypothesis. If the F-value is greater than the F-critical value, the model can reject the null hypothesis. Since the F-values in Figs. 7 and 9 are larger than the critical values, therefore, we can say that the geo location and user models are significant.

The F-distribution also shows that the hash tag time series failed to reject the null hypothesis because based

on the illustration in Fig. 8, its f-value of 0.186 is less than the f-critical value of 3.100.

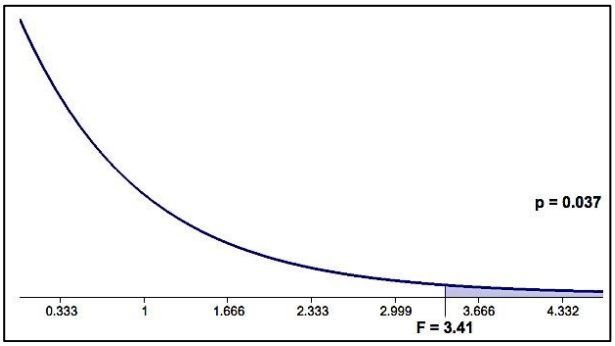

Fig.7. Geo Location Time Series with a Critical F-Value of **3.10006864.**

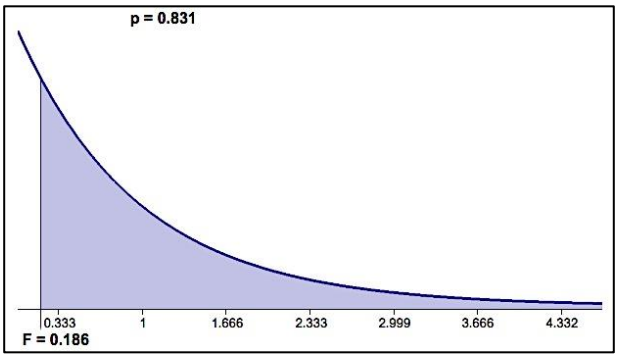

Fig.8. Hash Tag Time Series with a Critical F-Value of **3.10006864.**

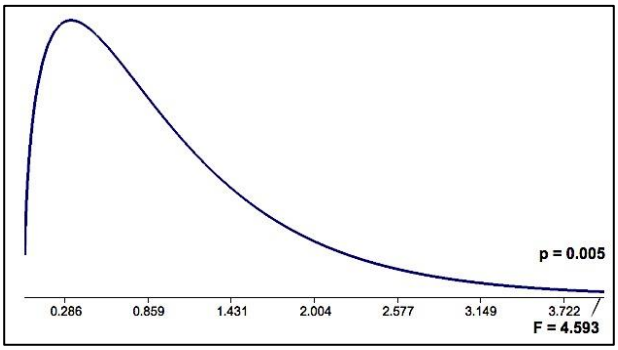

Fig.9. User Time Series with a Critical F-Value of **2.70699880.**

V. Results and Discussion

This paper investigated whether public sentiments or mood from a large collection of Twitter data was causative of the movement of the PSEi. Data were collected from June 2 – August 31, 2015, or 91 trading days. Among the three observed time series, tweets from geo location (Model 1) and users (Model 2) were show to have a predictive value for the stock movement.

The sentiment analysis classification results using the NLTK Naïve Bayes classifier produced 73% prediction accuracy as shown in Table 4. This bag of words method in feature extraction had the highest accuracy percentage compared with other methods that were used. Another model that was created using stop words and bigram collocations only produced 68% and 71% model accuracy. For that reason, the bag of words method was used as the classifier for sentiment analysis.

Table 4. Naïve Bayes Classifier Accuracy Check

Accuracy	Positive Precision	Positive Recall	Negative Precision	Negative Recall
0.728 (73%)	0.890	0.98	0.977	0.88

Precision measures the correctness of a classifier. The higher the precision, the more accurate the classifier – less false positives. Precision returns the fraction of test values that appear in the reference set. Recall on the other hand returns the fraction of reference values that appear in the test set. A high recall states that fewer reviews are identified and placed at the wrong label. The classifier that was used returned 89% and 97% positive and negative precision scores respectively, with a 98% positive recall and 88% negative recall.

To test the model, Table 5 shows the correlation coefficient (r) and coefficient of determination (r^2) of the models in a series of 91 days between the PSEi and Twitter time series. Although the r^2 goodness of fit appears to be strong, it doesn't imply the same results as the r.

Table 5. Correlation Coefficient and Coefficient of Determination for Models 1 and 2

Regression	r	r^2
Model 1	-0.11942	0.941167
Model 2	0.157607	0.945632

Geo location tweets for the past 2 days (lag-2) explained about 94% the regression variation of the predicted values in the time series of the PSEi. As with the user tweets for the past 3 days (lag-3) explained 95% of the regression variation of the predicted values in the time series of the PSEi. The remaining 6% and 5% are factors that were attributed to the error terms.

Figs. 10 and 11 show the scatterplots for the two prediction models based on their correlation coefficient.

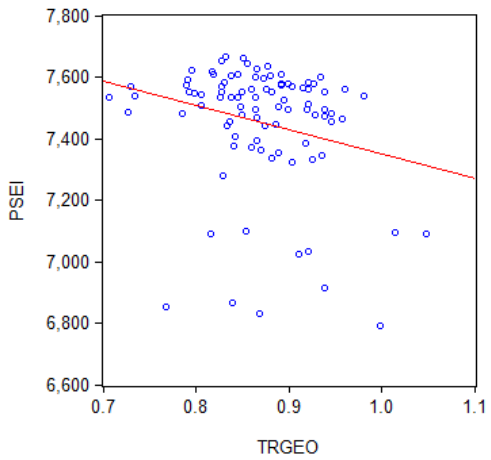

Fig.10. Scatterplot for Model 1 Correlation.

The (r) of the first model is -0.119420 which means that TRGeo and PSEi has a negative linear correlation. It implies that the relationship is if TRGeo increases (x), PSEi (y) decreases. Fig. 10 shows that whenever the polarity of the general mood of the public is high, there is a chance that the PSEi will decrease.

The second model has a correlation coefficient of (r) 0.157607, which means that TRUser and PSEi time series has a positive linear correlation.

Fig. 11 implies that the relationship is if TRUser increases (x), PSEi (y) also increases. Since the data from this model came from several news sources, we can say that whenever there is good news for the past three days, the future of the stock market is also affected, again, based on the Granger causality tests. Correlation check was used to see how good the models are in prediction, not for causation.

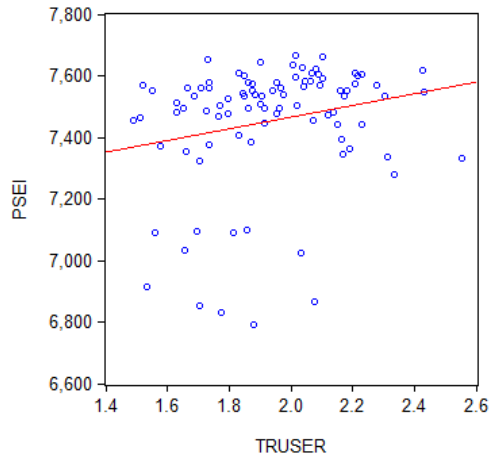

Fig.11. Scatterplot for Model 2 Correlation.

Based on the results in Table 6, we can reject two of our H_0 and go with the H_0, which are:

TRGeo does cause PSEi and
TRUser does cause PSEi

Table 6. P-values from Granger Causality Testing

Null Hypothesis	p-value
TRGeo does not cause PSEi	0.0376 (96.24%)
TRUser does not cause PSEi	0.0051 (99.49%)

Both models received highly significant p-values with 96.24% and 99.49% confidence level. P-values evaluate how well the sample data support that the H_0 is true.

If the P-value is:

$> .10 \rightarrow$ not significant
$\leq .10 \rightarrow$ marginally significant
$\leq .05 \rightarrow$ significant
$\leq .01 \rightarrow$ highly significant

High p-values indicate that the sample data is likely true with the H_0 and a low p-value indicates that the sample data is unlikely true with the H_0. Low p-values state that the time series provides enough evidence to reject the null hypothesis for the entire population. Thus, a low p-value indicated that the values obtained have occurred purely by chance, therefore, rejects the H_0.

Significance level is usually set at 0.05 or 5%, which means that experimental results that meet this significance level have, at most, a 5% chance of being the result of pure chance. Therefore, there's a 95% chance that the results were caused by the manipulation of experimental values.

As a result of the OLS regression, the researchers come up with these two prediction models based on the Granger causality results:

MODEL1: PSEi = C(1)*PSEI(-1)+C(2)*PSEI(-2)+ C(3)*TRGEO2(-1)+C(4)*TRGEO2(-2) + C(5)

Where:

C(1)	1.340883
C(2)	-0.375638
C(3)	198.1103
C(4)	-125.5645
C(5)	192.761

MODEL2: PSEI = C(1)*PSEI(-1) + C(2)*PSEI(-2) + C(3)*PSEI(-3) + C(4) *TRUSER2(-1) + C(5)* TRUSER2(-2) + C(6)*TRUSER2(-3) + C(7)

Where:

C(1)	1.18903
C(2)	-0.142392
C(3)	-0.067695
C(4)	-14.04438
C(5)	-37.32564
C(6)	-68.5666
C(7)	386.0128

Two models were tested based on the geo location and user time series. Figs. 12 and 13 show the time series of the actual UP and DOWN movement of the PSEi and the time series of the predicted values using the two prediction models. Both time series ware seen to meet in the same points and follow a similar direction with a 66% and 63% accuracy in results. The movement in the past values of the TRGeo and TRUser values predicts a similar rise and fall in the PSEi time series.

One interesting factor is the hash tag time series wasn't able to reject the null hypothesis because it didn't satisfy the p-value of <0.05. In the Granger causality test, the TRHash time series got a p-value of 0.8305 and an f-value of 0.186 in which both did not satisfy the significance level, therefore, failed to reject the H_0 and are not useful in predicting the PSEi. A possible reason is that the hash tag time series only has an average of 300 tweets per day compared to 8000 tweets for geo location and 1000 tweets for users. It shows that 300 tweets aren't enough to extract meaningful information.

The two prediction models based on the Granger causality analysis indicated that the geo location time

series with a 2-day lag and user time series with a 3-day lag are causative of the PSEi values with a high level of confidence. The models indicate that by looking at the past two or three days, we can be guided in the future movement of the stock market. Investors can use the models as another tool in helping them forecast the movement of the stock market in the domain of Twitter Sentiment Analysis.

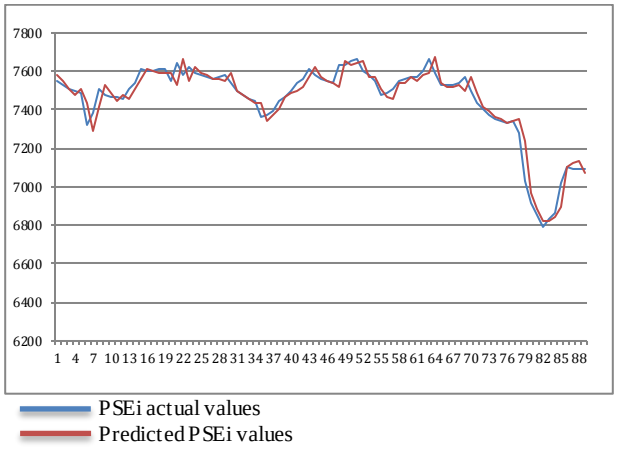

Fig.12. Time series of the actual PSEi values and predicted PSEi values using Model 1.

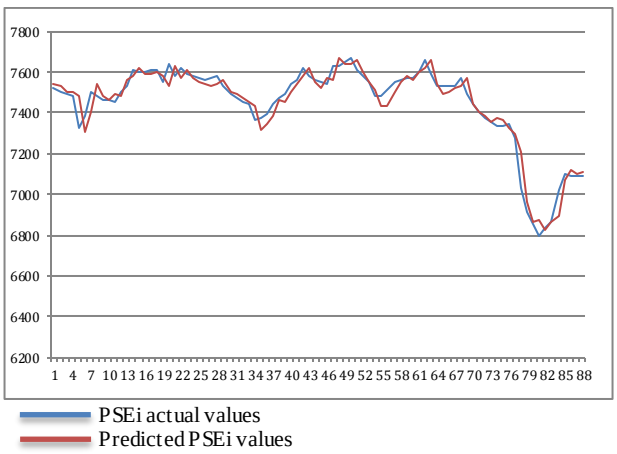

Fig.13. Time series of the actual PSEi values and predicted PSEi values using Model 2.

The lag depends on the AIC and each data were tested up to lags of 8. The first model received and AIC score of 7.928150 with a 2-day lag and the second model received a score of 10.57197 with a 3-day lag. The results show that the chatter coming from Twitter can be relevant in modeling the future performance of the stock market. However, the models do not predict the movement of specific stocks and only forecasts the stock market closing price index. Likewise, it does not predict the actual values of the PSEi, but provides a means of forecasting whether the PSEi movement will go UP or DOWN.

Of course, we must keep in mind that causation doesn't imply correlation; the models only estimate the predicted movement of the PSE closing stock index in the area of Granger causality and Sentiment Analysis.

This research investigated whether the general mood of the public and tweets from specific news sources is causative to the movement of the PSEi. It can be assumed that twitter sentiments have a predictive relation to the closing index of the PSE. Using the Naïve Bayes algorithm in assessing the sentiment of the tweets has proven, based on the p-values, to have a predictive power over the stock market.

VI. Conclusion and Future Work

The use of sentiment analysis and classification explored different approaches to incorporating its data with the stock market movement. After performing careful analysis, it can be therefore concluded that by using the Granger causality technique, wherein the past values of public tweets, jointly, can help explain the future values of the PSEi. Gathering about 800,000 tweets and collecting the closing values of the PSE within 3 months, two regression models were generated, based on the polarities of a tweet, in predicting the movement of the PSEi. However, only the tweets using the geo location and news sources provided us with the best results. Using hash tags in the sentiment analysis did not prove useful in rejecting the null hypothesis, thus, not helpful with the prediction.

This research investigated whether the general mood of the public and tweets from specific news sources is causative to the movement of the PSEi. It can be assumed that twitter sentiments have a predictive relation to the closing Index of the PSE. Using the Naïve Bayes algorithm in assessing the sentiment of the tweets have proven, based on the p-values, to have a predictive power over the stock market closing Index.

Overall, based on all the results of the tests and algorithms performed, Model 1, with the geo location time series, and Model 2, with the users time series, show causality with the PSEi movement. The hash tag time series failed to pass the causality tests and wasn't able to reject the null hypothesis.

The research provided two prediction models with different lags dependent variables, which proved to have a predictive power over the PSEi. It proposed a way to help investors in predicting the movement of the stock market by assessing the historical data the public tweets and the PSEi itself. This research does not predict the actual values of the PSEi but provided a way on how twitter sentiments can help tell whether the movement will go UP or DOWN.

For future work, here are some areas that can be improved on:

- A collection of more tweets per day, by using a larger time frame, may achieve greater results. Changes with the Twitter API posed several limitations in the data gathering of this research. Twitter limited the number of requests allowed per day and doesn't allow retrieval of tweets in the past.
- The sentiment analysis in this paper only identified two moods: Positive or Negative. Neutral

sentiments were not included, as they tend to be harder to identify because the classifier needs to determine the context of a sentence. This third dimension can be studied and compare the results with the ones presented here.

- Using different lexicon or corpora aside from the movie review lexicon used in this study may increase results in sentiment analysis. The MPQA, SentiWordNet, Opinion Lexicon and Profile of Mood States can be further explored.

REFERENCES

[1] Hong, H., Kubik, J. D., and Stein, J. C. (2004). Social Interaction and Stock-Market Participation. *The Journal of Finance*, 59(1), 137-163. doi: 10.3386/w8358

[2] Bollen, J., Mao, H., and Zeng, X. (2011). Twitter Mood Predicts the Stock Market. *Journal of Computational Science*, *2*(1), 1-8. doi: 10.1016/j.jocs.2010.12.007

[3] Chen, H., De, P., Hu, Y. J., and Hwang, B. H. (2014). Wisdom of Crowds: The Value of Stock Opinions Transmitted Through Social Media. *Review of Financial Studies*, 27(5), 1367-1403. doi: 10.1093/rfs/hhu001

[4] Dondio, P. (2013). Stock Market Prediction Without Sentiment Analysis: Using a Web-Traffic Based Classifier and User-Level Analysis. In *System Sciences (HICSS), 2013 46th Hawaii International Conference*. pp. 3137-3146. IEEE. doi: 10.1109/hicss.2013.498

[5] Yu, S., and Kak, S. (2012). A Survey of Prediction Using Social Media. *arXiv preprint arXiv:1203.1647*. unpublished.

[6] Oh, C., and Sheng, O. (2011). Investigating Predictive Power of Stock Micro Blog Sentiment in Forecasting Future Stock Price Directional Movement. *32nd ICIS Association for Information Systems*. 17.

[7] Oliveira, N., Cortez, P., and Areal, N. (2013). Some Experiments on Modeling Stock Market Behavior Using Investor Sentiment Analysis and Posting Volume from Twitter. In *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics*. p. 31. ACM. doi: 10.1145/2479787.2479811

[8] Mayfield, A. (2008). What is Social Media?. *iCrossing E-book*. Retrieved from http://ebooksoneverything.com/marketing/WhatisSocialMedia.pdf

[9] Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is Twitter, a Social Network or a News Media?. In *Proceedings of the 19th international conference on World wide web*. pp. 591-600. ACM. doi: 10.1145/1772690.1772751

[10] Asur, S., and Huberman, B. A. (2010). Predicting the Future with Social Media. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference*. Vol. 1, pp. 492-499. IEEE. doi: 10.1109/wi-iat.2010.63

[11] Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. (2010). Predicting Elections with Twitter: What 140 Characters Reveal About Political Sentiment. *ICWSM*, *10*, 178-185. doi: 10.2139/ssrn.1833192

[12] Bermingham, A., and Smeaton, A. F. (2011). On Using Twitter to Monitor Political Sentiment and Predict Election Results. *Sentiment Analysis where AI meets Psychology (SAAIP) Workshop at the International Joint Conference for Natural Language Processing (IJCNLP)*. 2-4.

[13] Bollen, J., Mao, H., and Pepe, A. (2011). Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena. *ICWSM*, *11*, 450-453.

[14] Mittal, A., and Goel, A. (2012). Stock Prediction Using Twitter Sentiment Analysis. *Standford University, CS229 2011*. Retrieved from*:* http://cs229. stanford. edu/proj2011/GoelMittal-StockMarketPredictionUsing TwitterSentimentAnalysis. pdf.

[15] Zhang, X., Fuehres, H., and Gloor, P. A. (2011). Predicting Stock Market Indicators Through Twitter "I hope it is not as bad as I fear". *Procedia-Social and Behavioral Sciences*, *26*, 55-62. doi: 10.1016/j.sbspro.2011.10.562

[16] Rao, T., and Srivastava, S. (2012). Analyzing Stock Market Movements Using Twitter Sentiment Analysis. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)* pp. 119-123. IEEE Computer Society.

[17] Ding, T., Fang, V., and Zuo, D. (2013). Stock Market Prediction based on Time Series Data and Market Sentiment. Retrieved from: http://murphy.wot.eecs.northwestern.edu/~pzu918/EECS349/final_dZuo_tDing_vFang.pdf

[18] Liu, B. (2012). Sentiment Analysis and Opinion Mining. *Synthesis lectures on human language technologies*, *5*(1), 1-167. doi: 10.1007/978-3-642-19460-3_11

[19] Granger, C. W. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica: Journal of the Econometric Society*, 424-438. doi: 10.2307/1912791

[20] Mao, H., Counts, S., and Bollen, J. (2011). Predicting Financial Markets: Comparing Survey, News, Twitter and Search Engine Data. *arXiv preprint arXiv:1112.1051*. unpublished.

[21] Bahadori, M. T., and Liu, Y. (2013). An Examination of Practical Granger Causality Inference. In *Proceedings of the 2013 SIAM International Conference on Data Mining*. doi: 10.1137/1.9781611972832.52

[22] Liew, V. K. S. (2004). Which lag length selection criteria should we employ?. *Economics bulletin*, 3(33), 1-9.

[23] Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: Sentiment Classification using Machine Learning Techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing*. Association for Computational Linguistics. Vol.10, 79-86. doi: 10.3115/1118693.1118704

[24] Zaki, M.J. and Meira, W. Jr. (2014). Data Mining and Analysis. Fundamental Concepts and Algorithms. *New York, New York: Cambridge University Press*.

Authors' Profiles

Anthony R. Caliñgo was born in 1987 in Manila, Philippines. He graduated with Academic Excellence at Informatics International College with the degree of Bachelor of Science in Information Management (BSIM) in 2008. He is now completing his Master's Degree in Information Technology (MIT) at the Technological Institute of the Philippines – Quezon City, under the guidance of Dr. Ariel M. Sison. His research interests include data mining and social media analysis.

He previously worked as an IT Instructor at Informatics International College from 2008-2014. Currently, he is the Assistant Administrator of Prince n' Princess School and a

Board Member of Prince n' Princess Corp., Pasig City, Philippines.

Dr. Ariel M. Sison earned his Doctor of Information Technology (DIT) at the Technological Institute of the Philippines Quezon City in 2013 and graduated with Highest Honors. He took up his master's degree in Computer Science at De La Salle University Manila in 2006 and obtained BS Computer Science at Emilio Aguinaldo College Manila in 1994.

He is currently the Dean, School of Computer Studies, Emilio Aguinaldo College Manila. His research interests include Data Mining and Data Security.

Dr. Sison is a member of International Association of Engineers (IAENG), Philippine Society of IT Educators and Computing Society of the Philippines. Currently, he is a Technical Committee Member of International Academy, Research, and Industry Association (IARIA) for International Conference on Systems (ICONS).

Dr. Bartolome T. Tanguilig III took his Bachelor of Science in Computer Engineering in Pamantasan ng Lungsod ng Maynila, Philippines in 1991. He finished his Master's Degree in Computer Science from De La Salle University, Manila, Philippines in 1999, and his Doctor of Philosophy in Technology Management from Technological University of the Philippines, Manila in 2003. He is currently the Assistant Vice President for Academic Affairs and concurrent Dean of the Graduate Programs of the Technological Institute of the Philippines, Quezon City.

Dr. Tanguilig III is a member of the Commission on Higher Education (CHED) Technical Panel for IT Education (TPITE), the chair of the CHED Technical Committee for IT (TCIT), the founder of Junior Philippine ITE Researchers (JUPITER), Vice President – Luzon of the Philippine Society of IT Educators (PSITE), board member of the PCS Information and Computing Accreditation Board (PICAB), member of the Computing Society of the Philippines (CSP) and a program evaluator / accreditor of the Philippine Association of Colleges and Universities Commission on Accreditation (PACUCOA).

How to cite this paper: Anthony R. Caliñgo, Ariel M. Sison, Bartolome T. Tanguilig III, "Prediction Model of the Stock Market Index Using Twitter Sentiment Analysis", International Journal of Information Technology and Computer Science (IJITCS), Vol.8, No.10, pp.11-21, 2016. DOI: 10.5815/ijitcs.2016.10.02