Applied Computer Systems
ISSN 2255-8691 (online)
ISSN 2255-8683 (print)
May 2020, vol. 25, no. 1, pp. 33–42
https://doi.org/10.2478/acss-2020-0004
https://content.sciendo.com

Predicting Stock Market Price Movement Using Sentiment Analysis: Evidence From Ghana

Isaac Kofi Nti[1, 2*], Adebayo Felix Adekoya[3], Benjamin Asubam Weyori[4]
[1, 3, 4] *Department of Computer Science and Informatics, University of Energy and Natural Resources, Sunyani, Ghana*
[2] *Department of Computer Science, Sunyani Technical University, Sunyani, Ghana*

***Abstract* – Predicting the stock market remains a challenging task due to the numerous influencing factors such as investor sentiment, firm performance, economic factors and social media sentiments. However, the profitability and economic advantage associated with accurate prediction of stock price draw the interest of academicians, economic, and financial analyst into researching in this field. Despite the improvement in stock prediction accuracy, the literature argues that prediction accuracy can be further improved beyond its current measure by looking for newer information sources particularly on the Internet. Using web news, financial tweets posted on Twitter, Google trends and forum discussions, the current study examines the association between public sentiments and the predictability of future stock price movement using Artificial Neural Network (ANN). We experimented the proposed predictive framework with stock data obtained from the Ghana Stock Exchange (GSE) between January 2010 and September 2019, and predicted the future stock value for a time window of 1 day, 7 days, 30 days, 60 days, and 90 days. We observed an accuracy of (49.4–52.95 %) based on Google trends, (55.5–60.05 %) based on Twitter, (41.52–41.77 %) based on forum post, (50.43–55.81 %) based on web news and (70.66–77.12 %) based on a combined dataset. Thus, we recorded an increase in prediction accuracy as several stock-related data sources were combined as input to our prediction model. We also established a high level of direct association between stock market behaviour and social networking sites. Therefore, based on the study outcome, we advised that stock market investors could utilise the information from web financial news, tweet, forum discussion, and Google trends to effectively perceive the future stock price movement and design effective portfolio/investment plans.**

***Keywords* – Artificial Neural Network, financial text mining, Ghana Stock Exchange, natural language processing, sentimental analysis, stock market prediction.**

I. Introduction

Stock market trend prediction aims at estimating the future price of a stock to enable investors make informed decisions on their investments. Though its prediction is alleged to be challenging due to its extraordinarily volatile and stochastic nature, the pursuit of maximising return on investment (ROI) has made it an essential task for financial analysts, investors and researchers [1]–[4].

This associated ROI has led to several and continuous attempts on stock market trend prediction targeted at improving prediction accuracy. Currently, stock prediction methods can be clustered into two, fundamental analysis and technical analysis. The fundamental analysis assesses the future stock price based on its related business or company performance, while technical analysis evaluates the future stock price based on its previous and present price and volume on the stock market utilising technical indicators [5], [6].

Nevertheless, literature shows that about 66 % of previous studies on the stock market were based on historical stock prices, while 23 % were based on fundamental analysis and 11 % on both methods [5]. Conversely, Pagolu *et al*. [3] argue that a firm equity value depends not only on historical stock-price data, but also on the current events, news, and product announcements. Notwithstanding the unpredictability of news and opinions, behavioural finance argues that the primary indicators for effective market prediction could be perceived through news and opinions, using platforms such as Social Networking Sites (SNSs) (also called social media) to enhance prediction performance [7], [8]. Nti *et al*. [5] also pointed out that the valuable information hidden in the news and SNSs can effectively serve as indicators for estimating stock market price movement.

Besides, the global increase in the development and use of SNSs as a communication medium amongst investors in the stock market has grown faster and more convenient [9]. Thus, the views of investors, which can affect their investment verdict may be exaggerated and spread at a faster rate via SNSs, which might affect the stock market to an extent [9]–[11].

Investor sentiment, according to Noura *et al*. [12] is defined as the perspectives and beliefs of investors about discount rate and future cash flow, which are not supported by the key fundamentals. Accordingly, sentiment analysis is the technique for transforming unstructured textual data to structured data, views, and emotions to generate useful insight and knowledge using natural language processing techniques, data mining and computational linguistics [2], [6]. When the sentiments or emotions of investors are negative, or distrust on social media,

* Corresponding author's email: Ntious1@gmail.com

it might persuade stock prices to drop. Likewise, when positive might persuade stock prices to rise than neutral sentiment [13].

Therefore, the investor sentiment constitutes a vital factor in determining the quoted price in the financial market [2], [14]. Research also reveals that thousands of online users' decisions and choice are based on online reviews [15]–[18]. As such, the analysis of investors' views has become a significant factor in stock market decision-making in recent years [19]–[21]. Aside from the rise in research on sentiment analysis in stock market prediction, few challenges such as misspelling, shortcuts and information duplication in text data have been reported in the literature to associate the sentiment analysis.

Talib *et al*. [22] argues that the efficiency of text mining algorithm in stock market predictions is low, as a result of duplication of the same information on different web news sites. Similarly, Wang *et al*. [23] pointed out that public opinion mined from the web is short and usually mixed with several misspelling, shortcuts, emojis, advertisements, images, and unusual grammar construction. Hence, it contributes to the complexity in machine learning application. These challenges have raised some disparity in views among researchers on the effect of textual data influence on stock market trends.

The paper by Nguyen *et al*. [24] reported that some studies argued that social media sentiments had weak or no predictive power, while others argued that social media had strong predictive abilities [24]. Therefore, the use of social media sentiments for stock market price predictions is still an open issue for research. Moreover, our partial search of the literature [5] shows that a very high percentage of the existing studies on sentiments from SNSs and stock market movements concentrated on developed nations. While, investigating the impact of investor sentiments on the stock market price movement of developing economies has been rarely discussed.

Nevertheless, as stated by Agarwal *et al*. [25] investor sentiments from developing nations on the stock market require extra attention from researchers, as the irrationality of investor behaviour and the degree of inadequacy of stock markets vary across states.

Given the above discussions, the current study seeks to determine whether the public sentiment of the Ghanaian investor mined from SNSs correlates with stock market movement on the Ghana Stock Exchange (GSE). Additionally, the study attempts to find out to what extent does public views influence the stock market future price. We also adopted ANN to handle the prediction task (whether the future price would rise or fall) on real-world datasets from the GSE.

The primary contributions of the current study are as follows: (i) We proposed a novel combination of a sentiment-driven dataset (from web news, Google trend, forum post and Twitter) to predict stock market price movement; (ii) we adopted the cosine similarity measure to prevent duplication of text dataset mined from different SNSs, thus overcoming the duplication weakness of text mining techniques pointed out in [22]; (iii) we adopted a Multi-Layer Perceptron (MLP) ANN (MLP-ANN) to handle the prediction task (whether the future price would rise or fall) on real-world datasets from the GSE; empirical results indicated a moderate increase in prediction accuracy of our amalgamated dataset in comparison with different data sources.

The remaining section of the current study is categorised as follows: Section II presents a brief discussion on machine learning and pertinent studies. Section III presents the study dataset and techniques and tools adopted for this study. In Section IV, we present the experimental results and discussion. Section V presents the main conclusions of the study.

II. Related Literature

This section presents a brief discussion of machine learning and a review of the state-of-the-art studies, which probed into the association between the financial markets and dataset from SNSs.

A. Machine Learning (ML)

ML is the practice where a robot or computer software learns from knowledge (K_w) relating to some group of tasks (T_k) and evaluation metric (E_p) if E_p in a task (T) improves with (K_w) [26]. In order to deal with the unstable, unstructured, disorderly, and nonlinear time series dataset, different learning-based algorithms such as Support Vector Machines (SVM), Decision Tree (DT), and ANN are mostly used in stock market predictions. In this study, we adopted the ANN due to its excellent learning capability for solving classification, prediction and regression problems [5].

B. Related Works

The literature shows that investment decisions are not entirely rational. As a result, numerous studies tried to understand how the investor is influenced while making an investment decision.

Several sentiment analytical tools exist in the literature; however, they can be categorised into two groups, namely, machine learning and word count analysis methods [13]. In the word count techniques, dictionaries are used to identify sentiment (positive or negative) for every word and then sum words' sentiment together [13]. Fundamentally, the negative words are counted and given different weights based on their negativism. Likewise, the positive words, and the party with the highest score "wins". Among the available word count techniques, Loughran and McDonald's financial lexicon and Harvard-IV dictionary are most commonly used in stock market predictions.

In the ML approach, the commonly used techniques are classification algorithms such as Support Vector Classifier (SVC), Naive Bayes and Neural Network (NN) [5]. One main drawback with this approach is the time required for manually labelling the training dataset. Regardless of the technique, one adopts, text data for sentiment analyses are generally mined from the web, and the three most considered sources are search engine queries, web financial news, and tweets from Twitter. Accordingly, we categorised our review of previous studies into these three data sources.

C. Search Engine Queries and Stock Price Movement

The relationship between Google search trends and stock market trading volume and volatility was carried out in [27]. The study examined whether search queries on Google could explain the current stock price and predict future abnormal returns of the stock market trends. The study outcome pointed out that Google searches could not predict future abnormal returns. Instead, the increased in search queries on Google predicted the increase in trading volume and volatility. Thus, from the study outcome, it can be established that Google searches are more associated with future than recent trading activity.

Likewise, an investigation into the effect of economic uncertainty and investor attention on gold price changes and their volatility was undertaken in [28]. Interestingly, the study found that Google searches initiated from India were more related to the gold market than searches made from the US or other countries. Contrariwise, the study established that increased in Google search volume was related to gold price drops and increased volatility. The relationship, according to the paper, was not just a correlation, but also showed a high predictive power in both directions.

Also, a technique for predicting stock market movements with web search data using an automatic search term selection with an adaptive approach was proposed in [29]. The study affirmed the predictability of stock movement with web search data from search engines. Again, the impact of search queries made on China's search engine (Baidu) on the volatility of China's stock markets was examined in [30]. The study found that stock market volatility was highly predictable with significant accuracy based on the Baidu-Index, particularly during the eras of economic instabilities.

Notwithstanding the achievement in stock market price prediction based on search engine queries, Bijl *et al*. [31] argued that trading techniques founded on search engine queries were profitable beforehand, but not after transaction costs.

D. Web Financial News and Stock Price Movement

Financial web news is reported to offer unobstructed and in-depth knowledge of the market because it is written by financial analysts [23]; this has resulted in several studies analysing its influence on the stock market. We present a few of these works pertinent to this study.

A sentiment analysis of news for predicting stock price movement using SVM enhanced with Particle Swarm Optimization (PSO) technique was proposed in [32]. The study confirmed a correlation between web news and stock price movement. The paper also reported an improvement in accuracy of 59.15 % compared with 57.8 % obtained in [33] using a deep learning model to examine web news and stock market movement.

Likewise, the association between the fluctuations in the sentimental tone of the European Central Bank's (ECB) news and stock price movement was studied in [18]. The study found a substantial effect of news sentiment tone on both the volatility and the mean of stock returns. Furthermore, the paper observed that the association between news sentiments and stock market volatility increased in strength during financial crisis. However, the sparsity of stock related news pointed out in [23] was a significant drawback for prediction models based on financial news, since machine learning models depended partially on the volume of the input dataset.

In another study, the information theory and Random Matrix Theory (RMT) were used to analyse the association between the stock market and news from the New York Times. The news was computed and transformed into everyday polarity time series [34]. The outcome of the study showed that news did not only correlate with the stock market but also was a high predictor of stock movement. The outcome of the study supported behavioural finance as the contemporary economic paradigm.

E. Twitter and Stock Market Movement

A trust management framework for the stock market based on stock related data (tweets) from Twitter was proposed in [13]. The study aimed at examining the degree of association between abnormal stock returns and Twitter sentiment. The study outcome showed a positive correlation between the two. Similarly, a study by Alshahrani and Fong [35] investigated the effect of sentiment analysis of tweets on stock market movement using the Fuzzy decision platform. The authors achieved a recall of 69 % minimum and 96 % maximum. Likewise, a sentiment analysis framework for predicting Brazilian stock market movement based on three perspectives: the absolute number of tweet sentiments, tweet sentiments weighted by favourites and tweet sentiments weighted by retweets was presented in [36]. The paper established that the stock market was predictable, based on tweet sentiments. The study also showed that MLP outperformed other state-of-the-art classifiers such as support vector machine and Naive Bayes.

However, few issues were identified in the literature to be associated with tweet data. Thus, there is suspected evidence of users posting under multiple accounts to sway opinion, which can affect the emerging market with less Twitter users and tweets. Again, most stock tweets are in response to past moves. Therefore, using tweets alone as input features for stock market analysis raises a question on model accuracy.

III. Materials and Methods

Based on the foregone argument in Section II of this paper, we observed that a minimal number of studies concentrated on examining the effects of investor sentiment on the stock markets in developing economies. Moreover, as argued in [37], a sentiment analysis model based on public views from one country might not work for all cases (generalisation). Therefore, understanding, or analysing people's sentiments from the global perspective might seem to be complex because several factors largely influence sentiments. Chiefly among them is cultural beliefs and practices, differences in policies and regulatory frameworks among different nations, differences in levels of stock market sophistication, among others. Thus, sentiment analysis might be easier with countries or region which share cultural affinity or proximity than from a global

level. It is in this light; we decided to investigate the impacts of investors sentiments about the Ghanaian Stock Market.

Likewise, as indicated in [23], [25], the accuracy of stock market predictions has been enhanced significantly but can be further improved beyond its current measure by looking for newer information sources on the Internet. Nonetheless, most of the previous studies discussed in Section II relied on information from Google or Twitter or the Internet stock message boards, or web news, but not on the combination of all. Moreover, a high percentage of existing studies focused on improving the predictability of stock market return while decreasing its volatility.

Hence, unlike previous studies discussed in Section II, we propose a novel amalgamation of several unstructured data sources from SNSs to address deficiencies in a single data source. We mix public opinion from tweets, web news, Google trends and forum discussion as a single input for more precise and accurate stock price trend prediction.

Specifically, the current study attempts to answer the following research question: Can investor sentiment truly help predict stock price movements, and to what extent? To the best of our knowledge, this study is the first study to use SNS data in examining the relationship between investor sentiment and stock market volatility on the GSE.

We present the details of the materials and techniques used to achieve the objectives of the current study in the subsequent section.

A. Study Framework

Figure 1 represents the data-pipeline framework for the current study. The framework is broken into five different steps, namely, dataset description, dataset preprocessing, dataset integration, predictive model and performance evaluation criteria. We explain below specific function of each step.

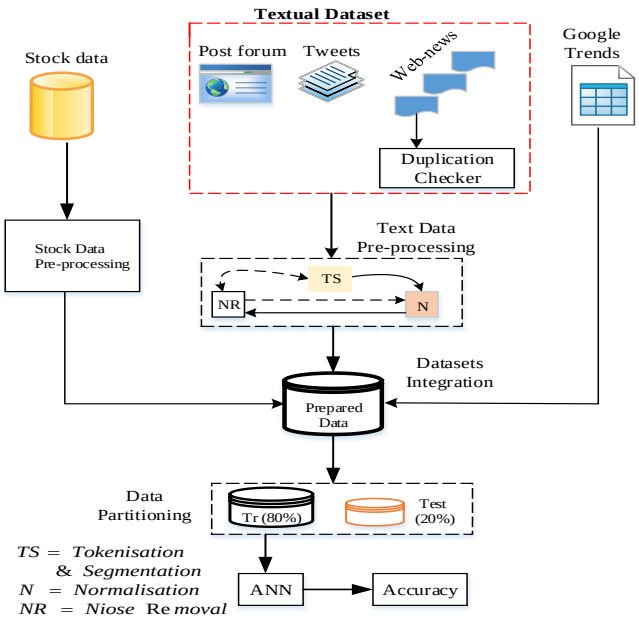

Fig. 1. Data pipeline of the study.

B. Dataset Description

Several methods are available for obtaining datasets; this includes an Application Programming Interface (API) provided by the organisation concerned, downloading the data from open source communities, and buying from data firms. We adopted the combination of these methods in our data gathering. The time-frame of study datasets was from January 2010 to September 2019.

In the current study, three datasets were used. The first was historical stock prices (D_{stock}) of three (3) companies (GCB, MTNGH and TOTAL) listed on GSE, which were downloaded from the official website of GSE (https://gse.com.gh). The dataset included "previous closing price", "opening price", "closing price" and "total shares traded".

The second was unstructured datasets which included tweets, web news, and post-forum. The tweet data were collected from Twitter, using a Twitter Search API Tweepy [38]. Besides, as done in several studies [6], [13], [14], [39], we used the dollar ($) sign to select stock market related tweets, since the dollar sign is usually used to tag stock tweets. We downloaded 2184 tweets. A total of 1581 financial news headlines relating to our companies of focus were collected from three well-known online news portals in Ghana using the Beautiful Soup API (i.e., myjoyonline.com, ghanaweb.com, and graphic.com.gh). We extracted our forum discussions with Pyglet API from (https://sikasem.org) a local platform that offers its users the opportunity to discuss financial and other matters concerning the GSE.

The news headlines and forum discussion downloaded were based on dates. However, unlike previous studies already discussed in Section II, we considered news spread among the public and people's comments on the news in addition to sentiments in news titles. We adopted the sentiment analyser [40] to obtain the collective sentiments from the forum messages concerning our companies of focus.

The third dataset was Google trends (D_{Gtrends}), a service provided by Google, which enables anyone to find out the volume of search on any topic. A total of 263 records were obtained from Google trends (https://www.google.com/trends) using Python and the pytrends API [41]. The search volumes were already scaled within 0 to 100, where 100 represented the highest search volume for any given period and 0 being the lowest volume. The Google trend search was restricted to only Ghana using keywords related to companies of focus in this study.

We normalised the dataset in the range of [0,1], by dividing each value with the maximum value of the data set for our ML model to function at a minimal computational time and enhanced efficiency.

C. Dataset Preprocessing

In this phase, we aimed at removing duplication of the same news presented on different websites using the cosine similarity check algorithm. The lexical and semantic similarity between news headlines were identified with the cosine similarity to minimise the duplication weakness in text mining algorithms pointed out in [22]. The cosine similarity measure was adopted

based on its simplicity and efficiency in estimating the similarity index between two vectors [42]. The similarity check algorithm tries to measure how 'close' two headlines are both in surface nearness (lexical similarity) and meaning (semantic similarity). This was accomplished by calculating the cosine distance between any two vectors (news headlines from different websites) as expressed in Eq. (1) [42].

$$Similarity(S) = \cos\theta = \left\{\frac{V_{ecA} \times V_{ecB}}{\|V_{ecA}\| \times \|V_{ecB}\|}\right\} = \left\{\frac{\sum_{i=1}^{n} V_{ecA_i} \times V_{ecB_i}}{\sqrt{\sum_{i=1}^{n} V_{ecA_i}^2} \times \sqrt{\sum_{i=1}^{n} V_{ecB_i}^2}}\right\}, \quad (1)$$

where V_{ecA} and V_{ecB} represent two sets of vectors, V_{ecAi} and V_{ecBi} = components of vector V_{ecA} and V_{ecB}, respectively. V_{ecA} is the Euclidean norm of the vector $V_{ecA} = (a_1, a_2, \ldots, a_i)$ defined as $\sqrt{(a_1^2 + a_2^2 + \ldots + a_i^2)}$. Likewise, $\| V_{ecB} \|$ represents the Euclidean norm of the vector V_{ecB}. The similarity measure (S) lies between [0 and 1]. Thus, two vectors (a_i & b_i) with the same orientation have a cosine similarity of 1, and two vectors (a_i & b_i) at 90° have a similarity of 0 [42]. Thus, $S = 1$ if ($a_i = b_i$) and $S = 0$ if ($a_i \neq b_i$).

The unstructured dataset (D_{text}) was passed through Tokenization and Segmentation (TS), Normalization (N), and Noise Removal (NR). This phase sliced the input texts up into smaller pieces, called tokens, while throwing away certain characters such as punctuation, symbols like #, @, /, %, URLs, extra spaces and stop words like "and," "a" and "the." We accomplished the text preprocessing of our dataset with the Natural Language Toolkit (NLTK) [40].

The sentiment in the tweet, web news and forum discussions were assessed in two dimensions, polarity and subjectivity as suggested in [43]. Polarity score was measured within the range [−1.0, 1.0], where 1.0 meant a positive statement and −1.0 meant a negative statement (see Fig. 2(A)). Subjective sentences usually refer to personal emotion, opinion or judgment; subjectivity is also a float which lies within [0.0, 1.0], while objective refers to truthful information. A score of 0.0 is considered to be very objective and 1.0 – very subjective (see Fig. 2(B)).

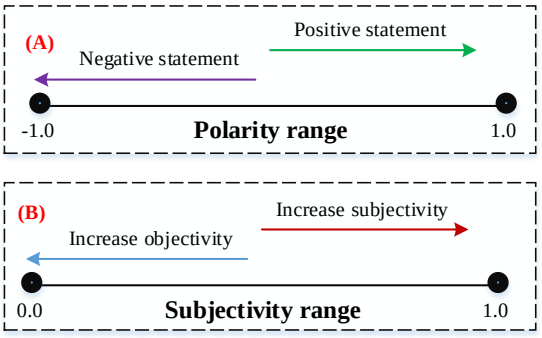

Fig. 2 Polarity and subjectivity ranges.

A total of twelve (12) features were extracted from the textual dataset (D_{text}). Table VI (Appendix A) shows the details of these features.

The stock historical price data (D_{Stock}) were used to examine if a stock price on (d) days ahead will fall or rise based on sentiments from SNSs. We substituted missing values in (D_{Stock}) with the average price (M_{Value}) of the day before and after the missing day, as defined in Eq. (2).

$$M_{value} = \frac{(Closing\ price_{d-1} + Closing\ Price_{d+1})}{2} \quad (2)$$

Thus, let $Closing\ Price_{d-1}$ = closing price of the stock the day before and $Closing\ Price_{d+1}$ = closing price of the stock a day after missing value's day. Like several studies [6], [13], [14], [23], [39], we aimed at stock returns. Hence, for a given date (i) and a given stock (S) and the target day (d), thus, if $d = 30$, then target date is $i + 30$ days. We estimated its percentage rise using Eq. (3), where $Closing\ Price_i$ = stock closing price for the i^{th} day and $Closing\ Price_{i+d}$ = stock closing price for the $i^{th} + d$. Thus, calculated stock returns $Return(S,t)$ reflected the change in stock price compared with the i^{th} day stock price. If $Return(S,t) > 0$, it implies a rise in next d day closing price, denoted as 1, and if $Return(S,t) < 0$, it represents a fall in the next (d) day closing price, denoted as 0 in this study.

$$Return(S,i) = \left(\frac{Closing\ Price_{i+d}}{Closing\ Price_i}\right) - 1 \quad (3)$$

D. Dataset Integration

At this stage, we integrated the three datasets (D_{text}), (D_{Gtrends}) and (D_{stock}) discussed above into a single file (DS) using Structured Query Language (SQL) server 2014 and Python. The date (d) variable was used as the reference point. Thus, the input dataset to the predictive model becomes a vector X, such that $X = \{x_{(0)}, x_{(1)}, x_{(2)}, \ldots, x_{(N-1)}, x_N\}$, where N represents the size of (DS). For every sample data $(X(t) \in \mathbb{R}^k)$ of (DS), where k stands for the number of features and t time stamp, $X = \{x_{(t)0}, x_{(t)1}, x_{(t)2}, \ldots x_{(t)k}\}$. The target value $(Return(S,i))$ denoted by y was represented as a sequence of labels.

Thus, $y = \{y_{(0)}, y_{(1)}, y_{(2)}, \ldots, y_{(N)}\}$, such that every element of $y \in \{0,1\}$. Hence for every input dataset (X), our predictive model makes a distinction between two classes, $y(t) = 1$ for positive and $y(t) = 0$ for negative. A classification of 1 denotes a "rise" and 0 a "fall" in stock returns ($Return_{ij}$) between a day before ($d - 1$) and day (d).

For training and testing of our proposed predictive framework (see Fig. 1), we divided the amalgamated dataset into two; 80 % for training and 20 % for testing based on literature [5].

E. Predictive Model

The Multi-Layer Perceptron (MLP) ANN algorithm was adopted for the current study due to its efficiency and effectiveness in predicting the financial market, as reported in several studies [4]–[6], [10], [44]. MLP is a network of interrelated components that accepts input, actuates, and then forwards it to the next layer. The MLP studies a function $f(.): R^D \rightarrow R^o$ by training on a dataset, where (D) represents the dimension of the input dataset, and o represents the number of dimensions for the output data.

Similar to several studies [6], [10], [45], we configured our MLP using the formulae ($2N$ +1), which is proven to be a best practice, where N = number of inputs. We implemented an MLP of configuration (6:33:33:33:1), the number of inputs = 16, having three hidden-layer (HL), HL1 and HL2 and HL3 (number of neurons per hidden layer = 33) for combined dataset. This configuration was altered for individual dataset based on configuration formula discussed. Bias/layer = 1, maximum iteration = 6000, optimizer = limited-memory BFGS (lbfgs), activation = ReLU in hidden layer and Sigmoid in output layer, learning rate of 0.001 for 25 epochs. The Back-Propagation algorithm was used in training the MLP model in this study. The experiments were conducted using the Python and Scikit-learn library (https://scikit- learn.org).

F. Performance Evaluation Criteria

Several statistical techniques are available for measuring the performance of machine learning models. However, as indicated in [6], [46] using only statistical metrics in evaluating a data science model in the financial field is not comprehensive. Hence, the current study adopted a combination of accuracy metrics (Specificity (true negative rate) as in Eq. (4), Sensitivity (true positive) as in Eq. (5), and Accuracy as in Eq. (6)) and closeness metrics (Root Mean Square Error (RMSE) as in Eq. (7), and Mean Absolute Percentage Error (MAPE) as in Eq. (8)), as defined in [5], [46].

$$Specitivity = \frac{TN}{TN + FP} \tag{4}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{5}$$

$$Accuracy = \frac{TN + TP}{FP + TP + TN + FN} \tag{6}$$

$$RMSE = \sqrt{\left(\frac{1}{N}\sum_{i=1}^{N}\left(t_i - y_i\right)^2\right)} \tag{7}$$

$$MAPE = \left(\frac{1}{N}\sum_{i=1}^{N}\left(\left|\frac{t_i - y_i}{t_i}\right|\right)\right), \tag{8}$$

where FN = incorrectly rejected (is false negative), TP = correctly identified (true positive), TN = correctly rejected (true negative), FP = incorrectly identified (false positive), y_i = predicted value by the model, t_i = actual value and N = total number of testing data.

IV. Results and Discussion

This section presents the results and discussion of the current study.

Tables I–V show the prediction performance (specificity, sensitivity, RMSE, MAPE and Accuracy) of the proposed model for different days (1, 7, 30, 60 and 90) ahead based on public sentiments from different data sources. Like several other studies [4], [6], [12]–[14], [18], [27], [28], [30], [32], [34], [35], this study affirms a degree of association between public sentiments and stock market trends, as shown in Tables I–V. Again, a strong positive correlation (0.9681) was observed between public sentiments and the volume of stock traded. Hence, it can be said that the association between public sentiments and stock market movement is not limited to only developed countries, but developing countries as well.

The results (see Tables I–V) show that the accuracy of the predictive model increases slightly with the combination of all data sources. This outcome suggests that the accuracy of stock market predictive models can be improved beyond its current measure by an amalgamation of newer information sources on the web affirming reportage in [23], [25]. Thus, adding more features from different stock related data sources can be one of the ways to enrich input features, which might enhance prediction accuracy.

Comparing the accuracy and error metrics recorded for different time window predictions, we observed that RMSE and MAE values kept decreasing, while prediction accuracy improved as the size of the window increased. This result agrees with findings in [36], [47] that sentiments analysis improves as prediction window sizes are increased. However, in some cases, the decrease in RMSE and MAE are high and low in other instance. We believe this happens for the reason that, as the prediction time window increases, it is possible for the MLP-ANN model to understand the prevalent sentiment in the SNSs datasets with better precision, as well as to see its influences on the future stock price movement.

Furthermore, the decrease in error metrics as the days ahead increase is affirmed by Khan *et al*. [48], who reported that the fundamental stock data were useful for medium- and long-term stock prediction than short-term. Accordingly, there is an indication that a sentimental event is more sensitive to new events.

Tables I–V show that in most of the experimental results, the combined data had lesser RMSE and MAPE values compared with the individual dataset.

TABLE I

1-DAY AHEAD PREDICTION OF STOCK PRICES BASED ON PUBLIC SENTIMENTS

Data sources	Specificity	Sensitivity	RMSE	MAPE	Accuracy (%)
Google trends	0.31	0.41	0.0276	2.7948	49.40
Tweets	0.42	0.49	0.0278	2.8238	55.50
Forum post	0.29	0.40	0.0521	3.3181	41.52
Web financial news	0.41	0.45	0.0495	3.1564	50.43
Combined data	0.51	0.69	**0.0251**	**2.5193**	70.66

TABLE II

7-DAY AHEAD PREDICTION OF STOCK PRICES BASED ON PUBLIC SENTIMENTS

Data sources	Specificity	Sensitivity	RMSE	MAPE	Accuracy (%)
Google trends	0.36	0.44	0.0240	2.4267	49.89
Tweets	0.45	0.50	0.0231	2.2924	56.30
Forum post	0.30	0.41	0.0441	2.7593	41.52
Web financial news	0.46	0.49	0.0405	2.5272	51.89
Combined data	0.59	0.71	0.0200	1.9468	73.69

TABLE III

30-DAY AHEAD PREDICTION OF STOCK PRICES BASED ON PUBLIC SENTIMENTS

Data sources	Specificity	Sensitivity	RMSE	MAPE	Accuracy (%)
Google trends	0.38	0.49	0.0173	1.7753	50.02
Tweets	0.49	0.52	0.0162	1.6431	56.98
Forum post	0.40	0.50	0.0308	1.8717	41.58
Web financial news	0.50	0.61	0.0274	1.6495	52.94
Combined data	0.61	0.79	0.0131	1.3444	75.02

TABLE IV

60-DAY AHEAD PREDICTION OF STOCK PRICES BASED ON PUBLIC SENTIMENTS

Data sources	Specificity	Sensitivity	RMSE	MAPE	Accuracy (%)
Google trends	0.49	0.51	0.0116	1.2024	52.87
Tweets	0.51	0.69	0.0107	1.1013	58.98
Forum post	0.41	0.49	0.0205	1.2385	41.77
Web financial news	0.42	0.56	0.0178	1.0651	54.05
Combined data	0.64	0.82	0.0087	0.8951	76.57

TABLE V

90-DAY AHEAD PREDICTION OF STOCK PRICES BASED ON PUBLIC SENTIMENTS

Data sources	Specificity	Sensitivity	RMSE	MAPE	Accuracy (%)
Google trends	0.50	0.61	0.0050	0.5123	52.95
Tweets	0.51	0.69	0.0047	0.4874	60.05
Forum post	0.41	0.49	1.0177	0.7647	41.77
Web financial news	0.43	0.62	0.9135	0.6663	55.81
Combined data	0.69	0.85	0.0034	0.3624	77.12

Combined data = Google trends + tweets + forum post + web financial news

On the other hand, sentiments from tweet had lower RMSE and MAPE values in some cases. Thus, the combined data outperformed an individual dataset, while tweets from Twitter outperformed Google trend, forum post and web financial news.

This result points out that ascertaining future stock prices does not only depend on historical stock prices and technical indicators, but it is also partially dependent on unstructured (fundamental) data. Therefore, predicting the stock market is perceivable through social networking sites and platforms.

Further analysis with F-score to examine the effect of the news spread (the number of sheared count), the number of comments made on a news article, positive sentiment and negative sentiments on stock price revealed that positive sentiments effect on stock price moving up was (32 %) as compared with negative sentiment (50 %). However, the

volume of stock traded daily was positively associated with positive sentiments (63 %) than negative (50 %). The result shows that the diffusion of negative or positive sentiments across the SNSs has a high potential in influencing stock market activities.

Again, we observed that the spread of news headlines had a better influence (64 %) on the volume of stock than the number of comments made on the news (62.2 %). At this point, the idea was to examine the influence of information diffusion over the Internet sources on investor behaviour. The obtained results affirm that the trading behaviour of Ghana investor is partially influenced by public news.

V. Conclusion

Research into stock market return has been on the increase lately. Different computational and soft computing techniques have been applied in this field to examine the predictability of the stock market based on sentiments from SNSs. Despite the achievement in prediction accuracy by previous studies, the literature argues that market prediction accuracy could be further improved beyond its current measure with newer information sources on the Internet as input features [23], [25].

We investigated the potential of public sentiment attitudes (positive vs. negative) and sentiment emotions (joy, sadness and more) extracted from web financial news, tweets, forum discussion and Google trends in predicting stock price movements, using MLP-ANN.

Our experimental setup with stock data (January 2010 to September 2019) of three (3) companies listed on the Ghana stock exchange shows that the stock market is predictable using public sentiments. The increase in accuracy recorded by the proposed model in predicting future stock price for 1 day, 7 days, 30 days, 60 days, and 90 days ahead (see Tables I–V) based on individual and combined datasets shows that the accuracy of stock prediction models can be significantly improved with stock related data amalgamation.

The outcome of this study is essential for policymakers, as they add up to the understanding of the transmission instrument of the monetary policy. Again, the Bank of Ghana (BoG) should be aware that each statement from their outfit could create unintended uncertainty in financial markets.

On the other hand, the limited size of the dataset obtained from Twitter shows that investors in developing countries such as Ghana hardly share their views on SNSs concerning market trends. Hence, it makes it insufficient to wholly depend on public sentiment from a single source to predict stock market movement in such countries.

Nonetheless, adding more stock related data sources enriched the stock market prediction accuracy, as observed in the study results. Hence, we trust that to solve this weakness and scarcity of fundamental data in developing markets, the use of an appropriate feature fusion technique is an exciting future direction to investigate. A solid combination of fundamental and technical approaches to market prediction in developing countries is also a motivating future direction to explore.

References

[1] A. E. Khedr, S. E. Salama, and N. Yaseen, "Predicting stock market behavior using data mining technique and news sentiment analysis," *International Journal of Intelligent Systems and Applications*, vol. 9, no. 7, pp. 22–30, Jul. 2017. https://doi.org/10.5815/ijisa.2017.07.03

[2] R. Ren, D. D. Wu, and T. Liu, "Forecasting stock market movement direction using sentiment analysis and support vector machine," *IEEE Systems Journal*, vol. 13, no. 1, pp. 760–770, Mar. 2019. https://doi.org/10.1109/JSYST.2018.2794462

[3] V. S. Pagolu, K. N. Reddy, G. Panda, and B. Majhi, "Sentiment analysis of Twitter data for predicting stock market movements," in *2016 International Conference on Signal Processing, Communication, Power and Embedded System*, 2017, pp. 1345–1350. https://doi.org/10.1109/SCOPES.2016.7955659

[4] F. Z. Xing, E. Cambria, and R. E. Welsch, "Intelligent asset allocation via market sentiment views," *IEEE Computational Intelligence Magazine*, vol. 13, no. 4, pp. 25–34, Nov. 2018. https://doi.org/10.1109/MCI.2018.2866727

[5] I. K. Nti, A. F. Adekoya, and B. A. Weyori, "A systematic review of fundamental and technical analysis of stock market predictions," *Artificial Intelligence Review*, vol. 53, no. 4, pp. 3007–3057, Apr. 2020. https://doi.org/10.1007/s10462-019-09754-z

[6] A. Picasso, S. Merello, Y. Ma, L. Oneto, and E. Cambria, "Technical analysis and sentiment embeddings for market trend prediction," *Expert Systems with Applications*, vol. 135, pp. 60–70, 2019. https://doi.org/10.1016/j.eswa.2019.06.014

[7] W. Chen, Y. Cai, K. Lai, and H. Xie, "A topic-based sentiment analysis model to predict stock market price movement using Weibo mood," *Web Intelligence*, vol. 14, no. 4, pp. 287–300, 2016. https://doi.org/10.3233/WEB-160345

[8] B. Li, K. C. C. Chan, C. Ou, and S. Ruifeng, "Discovering public sentiment in social media for predicting stock movement of publicly listed companies," *Information Systems*, vol. 69, pp. 81–92, Sep. 2017. https://doi.org/10.1016/j.is.2016.10.001

[9] K. Guo, Y. Sun, and X. Qian, "Can investor sentiment be used to predict the stock price? Dynamic analysis based on China stock market," *Physica A: Statistical Mechanics and its Applications*, vol. 469, pp. 390–396, 2017. https://doi.org/10.1016/j.physa.2016.11.114

[10] A. Pathak and N. P. Shetty, "Indian stock market prediction using machine learning and sentiment analysis," in *4th International Conference on Computational Intelligence in Data Mining*, 2019, pp. 595–603. https://doi.org/10.1007/978-981-10-8055-5_53

[11] S. N. Balaji, P. V. Paul, and R. Saravanan, "Survey on sentiment analysis based stock prediction using big data analytics," in *2017 Innovations in Power and Advanced Computing Technologies*, 2017, pp. 1–5. https://doi.org/10.1109/IPACT.2017.8244943

[12] N. Metawa, M. K. Hassan, S. Metawa, and M. F. Safa, "Impact of behavioral factors on investors' financial decisions: case of the Egyptian stock market," *International Journal of Islamic and Middle Eastern Finance and Management*, vol. 12, no. 1, pp. 30–55, 2019. https://doi.org/10.1108/IMEFM-12-2017-0333

[13] Y. Ruan, A. Durresi, and L. Alfantoukh, "Using Twitter trust network for stock market analysis," *Knowledge-Based Systems*, vol. 145, pp. 207–218, 2018. https://doi.org/10.1016/j.knosys.2018.01.016

[14] T. T. P. Souza and T. Aste, "Predicting future stock market structure by combining social and financial network information," *Physica A: Statistical Mechanics and its Applications*, vol. 535, 122343, 2019. https://doi.org/10.1016/j.physa.2019.122343

[15] D. M. E. D. M. Hussein, "A survey on sentiment analysis challenges," *Journal of King Saud University - Engineering Sciences*, vol. 30, no. 4, pp. 330–338, Oct. 2018. https://doi.org/10.1016/j.jksues.2016.04.002

[16] A. Bhardwaj, Y. Narayan, Vanraj, Pawan, and M. Dutta, "Sentiment analysis for Indian stock market prediction using Sensex and Nifty," in *4th International Conference on Eco-friendly Computing and Communication Systems*, 2015, pp. 85–91. https://doi.org/10.1016/j.procs.2015.10.043

[17] G. Ranco, D. Aleksovski, G. Caldarelli, M. Grčar, and I. Mozetič, "The effects of Twitter sentiment on stock price returns," *PLoS ONE*, vol. 10, no. 9, e0138441, 2015. https://doi.org/10.1371/journal.pone.0138441

[18] N. Apergis and I. Pragidis, "Stock price reactions to wire news from the European Central Bank: Evidence from changes in the sentiment tone and international market indexes," *Inter. Adv. in Economic Research*, vol. 25, no. 1, pp. 91–112, 2019. https://doi.org/10.1007/s11294-019-09721-y

[19] S. Poria, E. Cambria, and A. Gelbukh, "Aspect extraction for opinion mining with a deep convolutional neural network," *Knowledge-Based Systems*, vol. 108, pp. 42–49, 2016. https://doi.org/10.1016/j.knosys.2016.06.009
[20] M. V. Mäntylä, D. Graziotin, and M. Kuutila, "The evolution of sentiment analysis—A review of research topics, venues, and top cited papers," *Computer Science Review*, vol. 27, pp. 16–32, 2018. https://doi.org/10.1016/j.cosrev.2017.10.002
[21] S. Merello, A. P. Ratto, L. Oneto, and E. Cambria, "Predicting Future Market Trends: Which Is the Optimal Window?" in *INNS Big Data and Deep Learning Conference*, 2020. https://doi.org/10.1007/978-3-030-16841-4_19
[22] R. Talib, K. M. Hanif, S. Ayesha, and F. Fatima, "Text mining: Techniques, applications and issues," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 11, pp. 414–418, 2016. https://doi.org/10.14569/IJACSA.2016.071153
[23] Y. Wang, Q. Li, Z. Huang, and J. Li, "EAN: Event attention network for stock price trend prediction based on sentimental embedding," in *10th ACM Conference on Web Science*, 2019, pp. 311–320. https://doi.org/10.1145/3292522.3326014
[24] T. H. Nguyen, K. Shirai, and J. Velcin, "Sentiment analysis on social media for stock movement prediction," *Expert Systems with Applications*, vol. 42, no. 24, pp. 9603–9611, 2015. https://doi.org/10.1016/j.eswa.2015.07.052
[25] S. Agarwal, S. Kumar, and U. Goel, "Stock market response to information diffusion through internet sources: A literature review," *International Journal of Information Management*, vol. 45, pp. 118-131, Apr. 2019. https://doi.org/10.1016/j.ijinfomgt.2018.11.002
[26] T. Mitchell, *Machine Learning*, 1st Edition. McGraw Hill, 1997.
[27] N. Kim, K. Lučivjanská, P. Molnár, R. Villa, "Google searches and stock market activity: Evidence from Norway," *Finance Research Letters*, vol. 28, pp. 208–220, Mar. 2019. https://doi.org/10.1016/j.frl.2018.05.003
[28] J. Ho and L. H. Kristiansen, "Can Google Trends predict gold returns and its implied volatility?" Master's thesis, University of Stavanger, Norway, 2019.
[29] X. Zhong and M. Raghib, "Revisiting the use of web search data for stock market movements," *Scientific Reports*, vol. 9, 13511, 2019. https://doi.org/10.1038/s41598-019-50131-1
[30] J. Fang, G. Gozgor, C.-K. M. Lau, and Z. Lu, "The impact of Baidu index sentiment on the volatility of China's stock markets," *Finance Research Letters*, vol. 32, 101099, Jan. 2020. https://doi.org/10.1016/j.frl.2019.01.011
[31] L. Bijl, G. Kringhaug, P. Molnar, and E. Sandvik, "Google searches and stock returns," *International Review of Financial Analysis*, vol. 45, pp. 150–156, May 2016. https://doi.org/10.1016/j.irfa.2016.03.015
[32] R. Chiong, M. T. P. Adam, Z. Fan, B. Lutz, Z. Hu, and D. Neumann, "A sentiment analysis-based machine learning approach for financial market prediction via news disclosures," in *2018 Genetic and Evolutionary Computation Conference Companion*, 2018, pp. 278–279. https://doi.org/10.1145/3205651.3205682
[33] M. Kraus and S. Feuerriegel, "Decision support from financial disclosures with deep neural networks and transfer learning," *Decision Support Systems*, vol. 104, pp. 38–48, Dec. 2017. https://doi.org/10.1016/j.dss.2017.10.001
[34] A. García-Medina, L. Sandoval, E. U. Bañuelos, and A. M. Martínez-Argüello, "Correlations and flow of information between The New York Times and stock markets," *Physica A: Statistical Mechanics and its Applications*, vol. 502, pp. 403-415, 2018. https://doi.org/10.1016/j.physa.2018.02.154
[35] A. Alshahrani Hasan and A. C. Fong, "Sentiment analysis based fuzzy decision platform for the Saudi stock market," in *2018 IEEE International Conference on Electro/Information Technology*, 2018, pp. 23–29. https://doi.org/10.1109/EIT.2018.8500292
[36] A. E. O. Carosia, G. P. Coelho, and A. E. A. Silva, "Analyzing the Brazilian financial market through Portuguese sentiment analysis in social media," *Applied Artificial Intelligence*, vol. 34, no. 1, pp. 1–19, 2019. https://doi.org/10.1080/08839514.2019.1673037
[37] K. M. Swamy, "Sentiment Analysis with Tensorflow – TensorFlow and Deep Learning Singapore," 2017. [Online]. Available: https://engineers.sg/video/sentiment-analysis-with-tensorflow-tensorflow-and-deep-learning-singapore--1742.
[38] J. Roesslein, "Tweepy Documentation." [Online]. Available: http://docs.tweepy.org/en/latest/.
[39] R. Batra and S. M. Daudpota, "Integrating StockTwits with sentiment analysis for better prediction of stock price movement," in *2018 International Conference on Computing, Mathematics and Engineering Technologies*, 2018, pp. 1–5. https://doi.org/10.1109/ICOMET.2018.8346382
[40] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. O'Reilly Media Inc., 2009.
[41] J. Hogue and B. DeWilde, "Pytrends." [Online]. Available: https://pypi.org/project/pytrends/.
[42] B. Li and L. Han, "Distance weighted cosine similarity measure for text classification," in *14th International Conference on Intelligent Data Engineering and Automated Learning*, 2013, pp. 611–618. https://doi.org/10.1007/978-3-642-41278-3_74
[43] K. Ravi and V. Ravi, "A survey on opinion mining and sentiment analysis: Tasks, approaches and applications," *Knowledge-Based Systems*, vol. 89, pp. 14–46, Nov. 2015. https://doi.org/10.1016/j.knosys.2015.06.015
[44] S. Agrawal, D. Thakkar, D. Soni, K. Bhimani, and C. Patel, "Stock market prediction using machine learning techniques," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 5, no. 2, pp. 1099–1103, Mar.–Apr. 2019. https://doi.org/10.32628/CSEIT1952296
[45] B. W. Wanjawa, "Predicting future Shanghai stock market price using ANN in the period 21-Sep-2016 to 11-Oct-2016," 2016. [Online]. Available: https://arxiv.org/abs/1609.05394
[46] F. Z. Xing, E. Cambria, and R. E. Welsch, "Natural language based financial forecasting: a survey," *Artificial Intelligence Review*, vol. 50, no. 1, pp. 49–73, 2018. https://doi.org/10.1007/s10462-017-9588-9
[47] S. Dey, Y. Kumar, S. Saha, and S. Basak, "Forecasting to classification: Predicting the direction of stock market price using xtreme gradient boosting," 2016.
[48] H. Z. Khan, S. T. Alin, and A. Hussain, "Price prediction of share market using artificial neural network (ANN)," *International Journal of Computer Applications*, vol. 22, no. 2, pp. 42–47, May 2011. https://doi.org/10.5120/2552-3497

Isaac Kofi Nti holds HND in Electrical & Electronic Engineering from Sunyani Technical University, B. Sc. in Computer Science from Catholic University College, M. Sc. in Information Technology from Kwame Nkrumah University of Science and Technology. Mr Nti is a Lecturer at the Department of Computer Science, Sunyani Technical University, Sunyani, Ghana and currently is a Ph. D. candidate at the Department of Computer Science and Informatics, the University of Energy and Natural Resources Sunyani, Ghana. His research interests include artificial intelligence, energy system modelling, and intelligent information systems and social and sustainable computing, business analytics and data privacy and security.
E-mail: ntious1@gmail.com
ORCID iD: https://orcid.org/0000-0001-9257-4295

Adebayo Felix Adekoya holds B. Sc. (1994), M. Sc. (2002), and Ph. D. (2010) in Computer Science, an MBA in Accounting & Finance (1998), and a Postgraduate Diploma in Teacher Education (2004). He has put in about twenty-five (25) years of experience as a lecturer, researcher and administrator at the higher educational institution levels in Nigeria and Ghana. A. F. Adekoya is an Associate Professor of Computer Science and currently serves as the Dean, School of Sciences, University of Energy and Natural Resources, Sunyani, Ghana. His research interests include artificial intelligence, business & knowledge engineering, intelligent information systems, and social and sustainable computing.
E-mail: adebayo.adekoya@uenr.edu.gh
ORCID iD: https://orcid.org/0000-0002-5029-2393

Benjamin Asubam Weyori received his Ph. D. and M. Phil. in Computer Engineering from the Kwame Nkrumah University of Science and Technology (KNUST), Ghana in 2016 and 2011, respectively. He obtained his Bachelor of Science in Computer Science from the University for Development Studies (UDS), Tamale, Ghana in 2006. He is currently a Senior Lecturer and the Acting Head of the Department of Computer Science and Informatics, the University of Energy and Natural Resources (UENR) in Ghana. His main research interest includes artificial intelligence, computer visions (image processing), machine learning and web engineering.
E-mail: benjamin.weyori@uenr.edu.gh
ORCID iD: https://orcid.org/0000-0001-5422-4251

APPENDIX A

TABLE VI

INPUT FEATURES

Features	**Description**
Stock Data (D_{stock})	
Previous closing price	The closing price of the stock a day before
Opening price	The opening price of a stock on a day
Closing price	The closing price of a stock on a day
Total shares traded	The total number of shares traded on a day
Unstructured Data (D_{text})	
Tweets from Twitter	
ID	A unique ID of the tweet
Tweet Sentiment	The sentiment of the tweet
a. Subjectivity	The separated subjectivity from the tweet
b. Polarity	The separated polarity from the tweet
Favourite count	Number of favourites per tweet
Retweet count	Total number of retweets
Possible sensitive	The sensitivity of the tweet (Boolean true/false)
Financial web news	
News Sentiment	Sentiment in news
a. News Subjectivity	Separated subjectivity from news sentiments
b. News Polarity	Separated polarity from news sentiments
Shared	Number of sheared counts
Comments	Total number of comments on the news by the public
Forum Discussions	
Forum Sentiment	Sentiment in forum discussions
a. Forum Subjectivity	Separated subjectivity from forum sentiments
b. Forum Polarity	Separated polarity from forum sentiments
Forum Comments	Total number of comments on a topic posted on a forum
Google Trends (G_{trends})	
Google Trend index	Total number of trend counts