



Developing a supervised learning-based social media business sentiment index

Hyeonseo Lee¹ · Nakyeong Lee¹ · Harim Seo¹ · Min Song¹ 

Published online: 10 January 2019

© Springer Science+Business Media, LLC, part of Springer Nature 2019

Abstract

The fast-growing digital data generation leads to the emergence of the era of big data, which become particularly more valuable because approximately 70% of the collected data in the world comes from social media. Thus, the investigation of online social network services is of paramount importance. In this paper, we use the sentiment analysis, which detects attitudes and emotions toward issues of society posted in social media, to understand the actual economic situation. To this end, two steps are suggested. In the first step, after training the sentiment classifiers with several big data sources of social media datasets, we consider three types of feature sets: feature vector, sequence vector and a combination of dictionary-based feature and sequence vectors. Then, the performance of six classifiers is assessed: MaxEnt-L1, C4.5 decision tree, SVM-kernel, Ada-boost, Naïve Bayes and MaxEnt. In the second step, we collect datasets that are relevant to several economic words that the public use to explicitly express their opinions. Finally, we use a vector auto-regression analysis to confirm our hypothesis. The results show the statistically significant relationship between public sentiment and economic performance. That is, “*depression*” and “*unemployment*” lead to KOSPI. Also, it shows that the extracted keywords from the sentiment analysis, such as “*price*,” “*year-end-tax*” and “*budget deficit*,” cause the exchange rates.

Keywords Sentiment analysis · Social media · Machine learning · Supervised learning

✉ Min Song
min.song@yonsei.ac.kr

Hyeonseo Lee
sunnny92@naver.com

Nakyeong Lee
nakylee@yonsei.ac.kr

Harim Seo
seohr415@yonsei.ac.kr

¹ Library and Information Science, Yonsei University, Seoul, South Korea

1 Introduction

Many factors affect the economic performance and financial market. Particularly, social media networks and electronic devices are rapidly increasing factors. According to the reports by Pew Research Center [1] and Statistica [2], the number of social media users has been doubled since 2010 and is expected to increase from 0.97 billion users in 2010 to 2.95 billion users in 2020. Furthermore, the amount of data produced is expected to reach 44 zettabytes in 2020 [3]. The analytics using big data, which is collected from online social media, provides useful insights into practical applications [4]. Particularly, the opinions of the public toward a particular issue using social networking sites have a certain effect on our society. The sentiment analysis enables one to capture these emotions by identifying subjective contents as positive, neutral and negative emotions to certain social issues. Hence, the algorithm is frequently used in numerous areas such as financial market and marketing for sales forecasting. Zhang et al. [5] demonstrate that the social media sentiment provides statistically significant information about the stock price.

In this study, we use sentiment analysis to investigate the relationship between economic performance and public opinions. For this goal, we train sentiment classifiers with big data sources collected from various social media sites including news articles, twitter and blogs. Then, we consider three different feature sets, which include feature vector and sequence vector with positive and negative word dictionaries, emoticons and lexical properties of the sequence of words. Finally, we evaluate the performance of six classifiers: MaxEnt-L1, C4.5 decision tree, SVM-kernel, Ada-boost, Naïve Bayes and maximum entropy. The results show that MaxEnt-L1 has sustainably better performance than other classifiers. The next step is to predict the sentiments of the collected datasets with the trained classifiers and compare the sentiment scores with an economic index. Finally, we use a VAR analysis and Granger causality theory to investigate the causal relationship between the sentiment scores and the economic performance. The contribution of this paper is not to propose a new method but to deeply analyze the correlation between the economic value and the time series emotion value of the social data collected using specific keywords. The remainder of the paper consists of related works, methodology, economic results and conclusion.

2 Related work

2.1 Sentiment analysis

The sentiment analysis finds how sentiments are expressed in certain texts and whether favorable or unfavorable sentiments or opinions can be assigned in the texts. In other words, the sentiment analysis includes sentiment expressions, polarity and strength of the expression and the relationship among the subjects

[6]. As a topic of natural language processing (NLP) in the field of computer science, sentiment analysis has been studied in academic fields and industry. The primary purpose of applying sentiment analysis is to figure out how people feel about something. Furthermore, many companies, which have collected a lot of data on their customers and staffs, tend to use sentiment analyses to realize the reputation of their companies and make their business plan [7]. During the 2000s, sentiment analysis was developed in various areas. Pang and Lee [8] studied diverse theories and methodologies to approach sentiment analysis. Liu [9] mentioned that the major reason for the increase in sentiment analysis studies is the proliferation of social media. As noted in [8], advanced sentiment analysis methodology and applications are required to better understand customers.

Several works specifically proposed classifiers for sentiment analysis. According to [7], SA based on supervised learning is the most well-accepted approach for sentiment analysis. The common types of classifiers are Naïve Bayes classifier, decision tree, k-nearest neighbors, neural network, Support Vector Machine (SVM) and maximum entropy. Pang et al. [10] analyzed the performance of three classifiers (Naïve Bayes, maximum entropy, and SVM) on movie reviews with rating indicators, i.e., a number of stars, which served as a baseline. They also used star ratings as polarity signals in their training datasets. Nasukawa and Yi [6] and Wilson et al. [11] classified the contextual polarity of sentiment expressions. Particularly, [11] classified expressions about specific items using manually devised patterns to categorize the polarity. O'Hare et al. [12] reported that Multinomial Naïve Bayes better performed than the Support Vector Machine (SVM) on finance-related blogs. In their study, the classification on sentiments is for both ternary (i.e., positive, negative and neutral) and binary (i.e., positive and negative). In classifying the sentiment of Twitter messages, [13] concluded that machine learning algorithms such as Naïve Bayes, Support Vector Machine (SVM), and maximum entropy achieved a high accuracy (more than 80%) using the trained Twitter message with emoticons. They applied feature extractors that consisted of unigrams, bigrams, unigrams and bigrams, and unigrams with part of speech tags. However, sentiment classification is often perceived as having the domain-dependent problem because there are different sentiment expressions in different domains, and the same word can mean different sentiments. Therefore, [14] suggested a collaborative multi-domain sentiment classification approach to simultaneously train sentiment classifiers for multiple domains. Specifically, they disassembled the sentiment classifiers as a global one and a domain-specific one. Fernández et al. [15] proposed the Distributional Correspondence Indexing (DCI) method for domain adaptation in sentiment classification. The experiment of [15] shows that the DCI performs well in comparison with the latest technologies for cross-language and cross-domain sentiment classifications. In addition, DCI substantially reduces the computational cost and requires less human intervention. Also, [16] creates a *sentiment-related index (SRI)* to evaluate the association between different lexical elements in a specific domain with the help of domain-independent features as a bridge in order to reconcile the gap between different domains. Then, they suggest a new SRI-based cross-domain sentiment classification algorithm called *SentiRelated*, to analyze the sentiment polarity of short texts. Furthermore, [17] explains innovative approach to predicting the sentiment

of documents in multiple languages without translation through Latent Semantic Indexing (LSI) which is able to change over from multilingual corpus to a multilingual “concept space.” They invent and implement the experiments that examine the extent to which subjects and sentiment contribute individually to their classification accuracy. As a result, they try to straighten out the question of whether subjects and sentiment can be discerned sensibly.

2.2 Literature about using social media sentiment to business companies

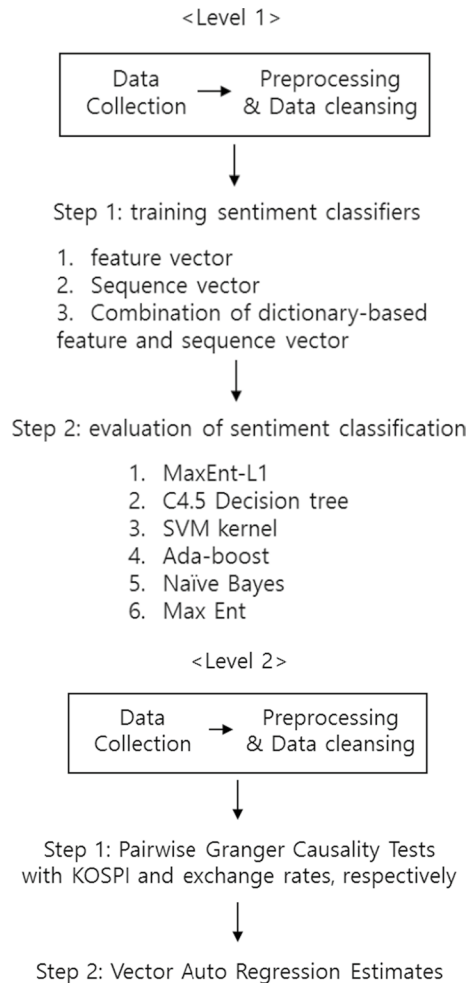
Various businesses embrace text and sentiment analysis and combine it into their processes because of its efficiency and accuracy. Manek et al. [18] suggested a statistical method using weight by Gini Index method with Support Vector Machine (SVM) for feature selection in sentiment analysis by using large movie review datasets. Many companies recently used social media data such as Facebook and Twitter more frequently to interact with customers. Culnan et al. [19] introduced the Fortune 500’s use of four of the most popular social media platforms: Twitter, Facebook, blogs and client-hosted forums. They showed case studies of three Fortune 100 corporations to demonstrate how they administered their respective networks of social media. Generally, most activities using social media are sales, customer care, advertising, marketing, product development and innovation [20]. He et al. [21] conducted an in-depth case study, which used text mining, to analyze the instruction text content on Facebook and Twitter sites of the three largest pizza chains: Pizza Hut, Domino’s pizza and Papa John’s Pizza. Yu et al. [22] examined the effect of social media and traditional media, their relative importance and correlations to short-term firm stock market performance. They exercised advanced sentiment analysis techniques beyond the number of mentions to analyze the overall sentiment of each media resource on a daily basis toward a particular company.

3 Methodology

Figure 1 shows an overview of this study. As shown, the process can be divided into two parts. The first part trains the sentiment classifiers and calculates the performance of their sentiment classification for six types of classifiers. The second part finds which economic indicators precede or follow with the sentiment score from each source of contents. After checking the rejected hypothesis at an alpha value of some variables, we ascertain whether there is an antecedent or an aftertaste among the variables. Then, a vector autoregressive analysis is used to find the time difference that the two variables show before and after. Therefore, the Granger causality test and vector autoregressive analysis were simultaneously performed in this study.

3.1 Collected data

To create a social media index that can be used to identify the public economy from social media data, we sought to index consumer responses to the welfare

Fig. 1 Research flow

economy based on a simple frequency of economic keywords. We collected 28 words of Twitter, blogs and news for each medium. In detail, in this study, we considered 73,229 news articles, 860,445 NAVER blogs and 9,749,893 tweets from Twitter from January 1, 2014, to October 31, 2015. We consider the periods between 2014 and 2015 because the Sewol ferry disaster occurred in 2014, and the Middle East respiratory syndrome (MERS) virus was running rampant during 2015. When we collected data, the terms of economic situation and event-related words were collected as query terms as shown in Table 1.

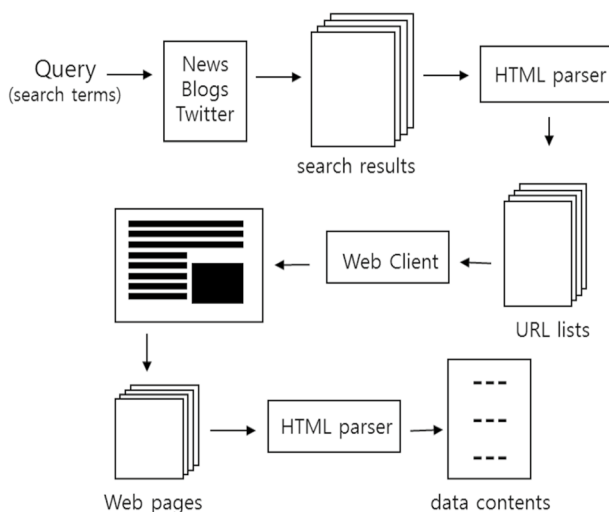
The data crawling process is shown in Fig. 2. When a specific query or search term is inputted, the search page results are collected. Using an HTML parser, the URL list is generated. With the Web client requests, web pages are gathered. Using the HTML parser, we can extract the data contents.

Table 1 Economic terms

boom, budget, budget deficit, central bank, consumer, corruption, depression, dollar, economy, export, finance, financial circles, growth, house, income, interest rate, major company, management, market, National Tax Service, oil price, policy, price, real estate, small business, unemployment, wage, year end tax

Economic terms in Korean

호황, 예산, 재정적자, 중앙은행, 소비자, 부패, 불황, 달러, 경제, 수출, 금융, 재계, 성장, 주택, 소득, 금리, 대기업, 경영, 시장, 국세청, 유가, 정책, 가격, 부동산, 중소기업, 실업, 임금, 연말정산

**Fig. 2** Data crawling process

Since the collected data are composed of a document unit, it must be cut into sentence units. We separated the sentence into tokens, which are semantic units, through the tokenization process, which removes whitespaces and measurement strings and divides the sentence into words. Lemmatization is a technique to group multiple forms of a single word into a single form. Stop word removal is the process of eliminating meaningless words such as articles, postpositions, prepositions and conjunctions. Morpheme analysis is the representation of the contents of words, phrases and paragraphs in the document as data that can be processed. It is possible to grasp the parts of the sentence morphemes and ultimately to understand the structure of sentences. This process is called part of speech (POS), which is a task of assigning parts of speech by processing words and assigning lexical categories to each word.

3.2 Selection of feature set

The following feature sets were fed into classifiers to predict the sentiments.

Positive and negative data were collected from various data sources in various manners, manually filtered and selected. We used positive and negative words that are circulating in the public. Word2vec was used to select candidate words as positive or negative and manually selected. The profanity data were added to the text by the Korea Creative Content Agency and divided into positive and negative.

- (1) Feature vector (including Korean positive word dictionary (11,461 words), Korean negative word dictionary (13,767 words), curse word dictionary (3863 words), positive emoticon dictionary (49), negative emoticon dictionary (52), Korean SentiWordNet (105,178 words)
- (2) Sequence vector (bag of words)–TokenSequence2FeatureSequence
- (3) Combine dictionary-based feature vector + bag of words

We compiled the training dataset for sentiment classifiers as follows. To make the classifier domain-neutral, first we collected 11,000 tweets using the query “Seoul Mayor Election.” Then, we collected 6000 news articles using the query “living cost and job.” Finally, we collected 2,450,000 movie reviews from NAVER. Because of the sheer volume of review data, we decided to use the movie ratings of customers. The scale of rate is 0–10; we considered ratings of 0–3 as negative, 4–7 as neutral and 8–10 as positive reviews. These datasets (except for movie review data) were independently reviewed by three evaluators. They labeled each text as negative, neutral or positive.

Among 17,000 data instances, the three judges agreed on 3230 data instances as positive, 5021 instances as neutral and 5410 instances as negative. The percentage of agreement is $(3230 + 5021 + 5410)/17,000 = 80.3\%$. Then, we used 13,661 data instances and 2.45 million movie reviews as the training data to learn the classifiers.

3.3 Machine learning algorithms

In this paper, we concentrate on selecting a correct classifier based on various feature set generation methods. Therefore, we apply six types of machine learning-based classification algorithms for evaluation: MaxEnt-L1, decision tree, SVM-kernel, Ada-boost, Naïve Bayes, and MaxEnt. MaxEnt, which is Max Entropy, is a probabilistic classifier and a type of exponential model that finds the probability distribution of maximum entropy [23]. MaxEnt is based on the principle of maximum entropy and can be applied to language detection, topic classification and sentiment analysis. Because we contribute to the performance of MaxEnt, we use MaxEnt-L1. According to [24], the MaxEnt model is a one-to-one relationship between subsets of variables that emerge from the parameterized factors of the model and subsets of variables to use in constraints. MaxEnt-L1, which adapts generalized expectation criteria for semi-supervised learning, has the flexibility to break out the one-to-one

relationship because the generalized expectation criteria are defined from the model that contains generalized expectation terms. In addition, generalized expectation criteria have many advantages such as the ease of use and simplicity [25]. The generalized expectation criteria do not need to have an additional process such as making an inverted index for pre-clustering unlabeled data. In this regard, we add MaxEnt-L1 to evaluate the measures. We also use the C4.5 decision tree classifier to approximate discrete valued functions using a decision tree; the C4.5 decision tree classifier is the most popular among inductive inference algorithms [26]. As another classifier, we use Ada-boost, which is fast and simple to program [27]. In addition, Ada-boost does not require prior knowledge about the base learner, so it can be combined with any other method to find the base classifiers. We also use Naïve Bayes, which is a probabilistic classifier based on Bayes theorem [28]. Using training data, Naïve Bayes predicts the category of documents using cue words that occur in the classified target document. Finally, we use the SVM [29], which can find a hyperplane divided by the maximal margin in the positive and negative subsets.

As evaluation measures of these classifier, there are four indicators: accuracy, recall, precision and F-measure. First, the accuracy represents the ratio of correct classification in the total classifications. Recall is the number of assigned proper classifications divided by the number of assigned total exact categories. Precision is the portion of correct categorizations in the total classification. The F-measure indicates the combination of precision and recall.

3.4 VAR analysis

In this section, we use a VAR analysis to identify the relationship between financial data such as KOSPI and the exchange rate among social media sentiments. Vector auto-regression (VAR) is a type of random process that enables one to detect the linear interdependencies among multiple time-series data. A VAR model describes how k variables evolve over time using their past values as follows.

A p th order VAR, which is denoted by VAR(p), is:

$$x_t = \alpha + A_1x_{t-1} + A_2x_{t-2} + \cdots + A_px_{t-p} + u_t$$

where x_{t-j} is the p th lag of x , α is a vector of constants, and u_t is an error term that satisfies $E(u_t) = 0$, $E(u_t, u_s) = \Omega$ and $E(u_t, u'_{t-p}) = 0$ where Ω is the covariance matrix of error terms.

The Korea Composite Stock Price index (KOSPI), which was first introduced in 1983 with the base value of 100, is computed from the prices of selected stocks using a weighted average. Levin and Zerovs [30] find that stock market predicts economic growth consistently. Hence, KOSPI can be used as an important indicator for economic activities.

3.5 Granger causality test

The fact that variable X is a Granger causality to variable Y implies that the fluctuation of the past X may affect the fluctuation of variable Y . Granger causality and

the precedence between variable X and variable Y can be determined by performing Granger causality test with different time lags. Granger causality test can be selected by inputting only two time series. The time difference or delaying time is set to 1, 2, 3, 4, 5 days, etc. The p value, which determines the hypothesis test result according to the delay time, can be used to estimate the relative Granger causality between the two variables. In this study, the alpha value (α) was set to 0.1, 0.05 and 0.01. After finding the rejected hypothesis at an alpha value of some variables, first we confirm whether there is an antecedent or an aftertaste among the variables. Then, a vector autoregressive analysis is used to find the time difference that the two variables show before and after. Therefore, Granger causality test and the vector autoregressive analysis were simultaneously performed in this study.

4 Results

4.1 Performance results of the sentiment classification

The performance results of sentiment classification are suggested in Table 2. Three types of feature sets have the highest $F-1$ in MaxEnt-L1: 0.7351, 0.7456, and 0.9296. When we use the vector feature set, the MaxEnt-L1 classifier indicates the highest accuracy (0.6787). In particular, when we combine the feature

Table 2 Sentiment classification performance results

	Accuracy	Recall	Precision	$F-1$
Feature set–vector				
MaxEnt-L1	0.6787 ± 0.0051	0.500	0.708	0.74
Decision tree	0.5096 ± 0.0036	0.513	0.597	0.38
SVM-kernel	0.4778 ± 0.0099	0.500	0.483	0.62
Ada-boost	0.6695 ± 0.0049	0.259	0.323	0.23
Naïve Bayes	0.6763 ± 0.0052	0.500	0.450	0.41
MaxEnt	0.5129 ± 0.0027	0.511	0.516	0.47
Feature set–sequence				
MaxEnt-L1	0.8929 ± 0.0168	0.663	0.746	0.75
Decision tree	0.6834 ± 0.0018	0.665	0.699	0.67
SVM-kernel	0.8942 ± 0.0226	0.639	0.719	0.70
Ada-boost	0.9153 ± 0.0143	0.644	0.753	0.67
Naïve Bayes	0.3789 ± 0.0221	0.500	0.448	0.49
MaxEnt	0.9091 ± 0.0145	0.646	0.74	0.65
Feature set–combined				
MaxEnt-L1	0.9353 ± 0.0076	0.931	0.930	0.93
Decision tree	0.6834 ± 0.0018	0.665	0.699	0.67
SVM-kernel	0.8590 ± 0.0188	0.747	0.812	0.80
Ada-boost	0.8942 ± 0.0226	0.639	0.719	0.70
Naïve Bayes	0.8751 ± 0.0122	0.620	0.714	0.62
MaxEnt	0.9556 ± 0.0071	0.500	0.903	0.87

vector and bag of words, recall, precision, and $F-1$ have the highest values in MaxEnt-L1. As a result, MaxEnt-L1 has better performance than five other classifiers.

4.2 VAR analysis

4.2.1 VAR analysis with KOSPI

The fact that variable X is a Granger causality to variable Y implies that the fluctuation of the KOSPI and economic-related keywords such as “boom,” “depression” and “unemployment” were selected to investigate the relationship between the financial market and the sentiment scores using a VAR analysis. The VAR model is known as a successful technique to predict interrelated time-dependent variables, structural inference and policy analysis. In this study, we consider four endogenous variables for the VAR analysis: KOSPI, “boom,” “depression” and “unemployment.” Furthermore, we use Granger causality test to identify the causal relationship between the KOSPI and four other keywords selected from social media.

Before Granger causality test is applied, it is necessary to determine the optimal lag length because Granger methodology is sensitive to the lag length. From the results of Akaike information criterion (AIC), the 5-lag length is selected as an appropriate lag structure for the variables. Granger causality test procedure involves estimating the following series of regressions. Each variable in this system depends on its own lags and the lags of other variables.

$$Z_t = c + \sum_{i=1}^5 A_i Z_{t-i} + \epsilon_t \quad (1)$$

where Z_t is an $n \times 1$ vector variable. The vector of variables in the VAR is $Z_t = [y_t \ b_t \ d_t \ u_t]^T$, which includes KOSPI (denoted by y), extracted keywords “boom,” “depression” and “unemployment,” which are denoted by b_t , d_t and u_t , respectively.

$E(\epsilon_t) = 0$, $E(\epsilon_t, \epsilon_s) = 0$ for $s \neq t$, and

$$E(\epsilon_t, \epsilon'_t) = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2 & \sigma_1\sigma_3 & \sigma_1\sigma_4 & \sigma_1\sigma_5 \\ \sigma_1\sigma_2 & \sigma_2^2 & \sigma_2\sigma_3 & \sigma_2\sigma_4 & \sigma_2\sigma_5 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sigma_1\sigma_5 & \sigma_2\sigma_5 & \sigma_3\sigma_5 & \sigma_4\sigma_5 & \sigma_5^2 \end{bmatrix}.$$

The coefficients $A_i = [\beta_{1i} \ \beta_{2i} \ \beta_{3i} \ \beta_{4i}]$ are constants to be estimated. The test results can be obtained from Eq. (1).

$$(i) \quad H_{o(1)} : \beta_{21} = \beta_{22} = \dots \beta_{25} = 0.$$

$$(ii) \quad H_{o(2)} : \beta_{31} = \beta_{32} = \dots \beta_{35} = 0.$$

$$(iii) \quad H_{o(3)} : \beta_{41} = \beta_{42} = \dots \beta_{45} = 0.$$

Table 3 Pair-wise Granger causality tests

Null hypothesis	F-Stat
BOOM does not Granger Cause KOSPI	1.25
KOSPI does not Granger Cause BOOM	2.20**
DEPR. does not Granger Cause KOSPI_	6.86***
KOSPI does not Granger Cause DEPR.	1.44
UNEMP does not Granger Cause KOSPI_	4.96***
KOSPI does not Granger Cause UNEMP	2.38**

Asterisks ** and *** correspond to 5% and 1% significance, respectively

The above hypotheses can be interpreted as follows: The test analyzes the null hypothesis that: (1) The keyword “boom” does not cause KOSPI, (2) “depression” does not cause KOSPI, and (3) “unemployment” does not cause KOSPI. Hence, the test results in Table 3 show that “depression” and “unemployment” lead to KOSPI, whereas KOSPI causes “boom” and “unemployment.” Consequently, there is a bi-directional causality in the short-run dynamics between KOSPI and “unemployment.” The results reveal uni-directional relationships between “depression” and KOSPI and between “unemployment” and KOSPI. If we reject the null hypothesis of (i), then we conclude that there is a causality from “boom” to KOSPI.

The outcome of Granger causality test to determine the interaction among KOSPI, “boom,” “depression” and “unemployment” for the specified period is shown in Table 3. The results show that both null hypotheses $\beta_{31} = \beta_{32} = \dots \beta_{35} = 0$ and $\beta_{31} = \beta_{32} = \dots \beta_{35} = 0$ are rejected. Consequently, “depression” and “unemployment” lead to KOSPI.

For each parameter estimate in Table 4, “boom” with lag 1 and lag 3 are statistically significant at the 10-percent level; “depression” with lag 2 and lag 4 are statistically significantly different from zero. Finally, “unemployment” at $t-1$ and $t-2$ have a statistically significant effect on the KOSPI. Hence, the selected keywords relating to economic terms such as “boom,” “depression” and “unemployment” with lags have a significant effect on the price of KOSPI. Furthermore, the coefficients of the KOSPI index with lag 3 are significantly different from zero.

Table 5 shows the results of AIC and BIC values that were used as a criterion for model selection. Given the results, we prefer the model with the lowest AIC or BIC value. Hence, we prefer the fifth lag with the lowest AIC or BIC.

4.2.2 VAR analysis with exchange rates

In this study, we consider four endogenous variables: exchange rates, “price,” “year-end-tax” and “budget deficit.” Given the Akaike information criterion (AIC), we choose lag 2 for the optimal lag length.

The outcome of Granger causality test to determine the interaction among the exchange rate, “price,” “year-end-tax” and “budget deficit” for the specified period is indicated in Table 6. The results present that the extracted keywords

Table 4 Vector auto-regression estimates

	KOSPI	BOOM	DEPR	UNEMP
KOSPI(− 1)	0.97*** (0.05)	− 0.01 (0.19)	0.17 (0.20)	− 0.07 (0.17)
KOSPI(− 2)	− 0.01 (0.07)	0.05 (0.26)	0.24 (0.27)	0.46* (0.24)
KOSPI(− 3)	− 0.09 (0.07)	− 0.46* (0.26)	− 0.51* (0.27)	− 0.68*** (0.24)
KOSPI(− 4)	0.13* (0.07)	0.57** (0.26)	− 0.02 (0.27)	0.06 (0.23)
KOSPI(− 5)	− 0.03 (0.05)	− 0.14 (0.18)	0.01 (0.19)	0.31* (0.17)
BOOM(− 1)	− 0.02* (0.01)	− 0.04 (0.05)	− 0.003 (0.05)	− 0.09** (0.04)
BOOM(− 2)	0.01 (0.01)	0.09* (0.05)	0.07 (0.05)	− 0.04 (0.04)
BOOM(− 3)	− 0.03** (0.01)	0.01 (0.05)	0.09* (0.05)	0.02 (0.05)
BOOM(− 4)	0.004 (0.01)	− 0.08 (0.05)	0.01 (0.05)	− 0.08* (0.05)
BOOM(− 5)	0.002 (0.01)	− 0.07 (0.05)	0.02 (0.05)	0.01 (0.05)
DEPR.(− 1)	− 0.001 (0.01)	0.06 (0.05)	0.09* (0.05)	− 0.07* (0.04)
DEPR.(− 2)	− 0.039** (0.01)	0.02 (0.05)	0.03 (0.05)	0.04 (0.04)
DEPR.(− 3)	0.02* (0.01)	− 0.05 (0.05)	− 0.04 (0.05)	0.02 (0.04)
DEPR.(− 4)	0.02 (0.01)	− 0.03 (0.05)	0.05 (0.05)	0.07* (0.04)
DEPR.(− 5)	− 0.05** (0.01)	0.05 (0.05)	0.06 (0.05)	0.03 (0.04)
UNEMP.(− 1)	0.03* (0.01)	− 0.06 (0.05)	0.13** (0.06)	0.23*** (0.05)
UNEMP.(− 2)	− 0.06*** (0.01)	− 0.09* (0.05)	− 0.05 (0.06)	0.002 (0.05)
UNEMP.(− 3)	0.01 (0.01)	− 0.07 (0.05)	− 0.07 (0.06)	0.05 (0.05)
UNEMP.(− 4)	− 0.02 (0.01)	0.07 (0.05)	0.22*** (0.06)	0.33*** (0.05)
UNEMP.(− 5)	0.01 (0.01)	0.14*** (0.06)	− 0.09 (0.06)	− 0.07 (0.05)
C	− 0.01 (0.01)	− 0.01 (0.05)	− 0.03 (0.05)	0.01 (0.04)
Likelihood	− 26.28	− 608.18	− 631.13	− 567.01

Table 4 (continued)

Standard errors are in (). The asterisks *, ** and *** correspond to 10%, 5% and 1% significance, respectively. DEPR and UNEMP represent DEPRESSION AND UNEMPLOYMENT, respectively
($-j$) indicates the j -period back observation of the data. For example, (-1) signifies the day before data

Table 5 AIC and BIC values

Lag	AIC	BIC
0	11.40967	11.44702
1	8.814046	9.000771*
2	8.773471	9.109575
3	8.786270	9.271754
4	8.680262	9.315125
5	8.672718*	9.456961
6	8.692697	9.626320
7	8.697282	9.780285
8	8.698914	9.931297

* Lag order selected by the criterion

Table 6 Pair-wise Granger causality tests

Null hypothesis	F -Stat
PRICE does not Granger Cause EXCHANGE_RATE	4.05**
EXCHANGE_RATE does not Granger Cause PRICE	1.79
YEAR_END_TAX does not Granger Cause EXCHANGE_RATE	2.59**
EXCHANGE_RATE does not Granger Cause YEAR_END_TAX	1.94
BUDGET_DEFICIT does not Granger Cause EXCHANGE_RATE	3.44**
EXCHANGE_RATE does not Granger Cause BUDGET_DEFICIT	1.21

The asterisks ** correspond to 5% significance

from the sentiment analysis, such as “*price*,” “*year-end-tax*” and “*budget deficit*,” cause the exchange rates.

As shown in Table 7, the estimated coefficients of “*price*” and “*year-end-tax*” with lag 2 are statistically significantly different from zero at least at the 10% level. The lagged value of exchange rates significantly affects the “*price*.” Therefore, Granger causality runs one-way from price, “*year-end tax*” and “*budget deficit*” to exchange rate (Table 8).

Regarding the VAR analysis of exchange rates, we prefer the second lag that minimizes both AIC and BIC values. Hence, we determine the second lag for the VAR analysis.

Table 7 Vector auto-regression model estimates

	EX. Rate	PRICE	YEAR_END TAX	BUDGET DEFICIT
EX.Rate(− 1)	0.09** (0.04)	− 0.19 (2.98)	3.64 (2.27)	− 0.01 (2.83)
EX.Rate(− 2)	0.06 (0.04)	− 5.62* (2.97)	2.15 (2.26)	4.30 (2.82)
PRICE(− 1)	0.00 (0.00)	0.1*** (0.05)	− 0.01 (0.04)	− 0.03 (0.05)
PRICE(− 2)	0.00*** (0.00)	0.05 (0.05)	0.06 (0.04)	− 0.07 (0.05)
YEAR END TAX(− 1)	0.00 (0.001)	− 0.01 (0.06)	0.47*** (0.05)	− 0.03 (0.06)
YEAR END TAX(− 2)	0.00* (0.00)	0.05 (0.06)	0.24*** (0.05)	0.05 (0.06)
BUDGET DEFICIT (− 1)	0.00*** (0.00)	− 0.08* (0.05)	0.028 (0.04)	0.15*** (0.05)
BUDGET DEFICIT (− 2)	− 0.00 (0.00)	− 0.04 (0.06)	− 0.02 (0.04)	0.15*** (0.05)
C	− 27*** (1.93)	− 187 (128.1)	186.42* (97.74)	138.0 (121.7)
Likelihood	1239.9	− 619.2	− 499.27	− 596.41

Standard errors are shown in (). The asterisks *, ** and *** correspond to 10%, 5% and 1% significance, respectively. Ex. Rate signifies the Exchange Rate

(− *j*) indicates the *j*-period back observation of the data

Table 8 AIC and BIC values

Lag	AIC	BIC
0	2.892807	2.930151
1	2.372808	2.559533*
2	2.302566*	2.638671
3	2.317637	2.803121
4	2.348656	2.983520
5	2.379977	3.164220
6	2.394989	3.328613
7	2.440544	3.523546
8	2.435519	3.667902

* Lag order selected by the criterion

5 Conclusion

On the economic side, sentiment analysis is a notably interesting field of research. In this study, we conducted experiments using six classifiers to analyze the sentiment

of the public in social media related to several economic words. We combined the machine learning method, statistical analysis and Korean economy. Then, we investigated the relation among the sentiments from three types of media (i.e., news, Twitter and blogs) and actual economic indicators such as KOSPI and exchange rates by applying Granger causality test and vector auto-regression model. We found whether the sentiment scores derived from large-scale datasets were correlated with the economic index over time. The results show that MaxEnt-L1 surpasses other classifiers that we expect. In addition, we used a VAR analysis to investigate the relationship between the sentiment of the public and the actual economic situation related to the economic theme. We confirm that the sentiment of the public shown in some economic words is actually related to the economic situation. In other words, analyzing the public sentiment can result in meaningful economic forecasts or useful information in the enterprise. In fact, a company that analyzes and uses the public sentiment through social media has a stronger effect on operations [12, 14]. Therefore, it is expected that companies will be able to see good effects if they recognize the importance of public sentiment analysis and apply it to their marketing, customer service and operation methods. In future research, we plan to show the public sensibility related to economic keywords and the effect on the actual economic situation by comparing the economic index with the more in-depth emotion of the public. In addition, the effect on the actual economic situation should be demonstrated instead of the public sensibility related to only few economic keywords by comparing the economic index with the more in-depth emotion of the public.

Acknowledgements This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2015S1A3A2046711).

References

1. Perrin A (2015) Social media usage. Pew research center, pp 52–68
2. Statista, Number of social network users worldwide from 2010 to 2021 (in billions). <https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/>
3. Jay Jacobs, CFA (2016) Social Media: Tech's Growth Industry. <https://www.globalxfunds.com/social-media-techs-growth-industry/>
4. Jin S, Lin W, Yin H, Yang S, Li A, Deng B (2015) Community structure mining in big data social media networks with MapReduce. *Clust Comput* 18(3):999–1010
5. Zhang G, Xu L, Xue Y (2017) Model and forecast stock market behavior integrating investor sentiment analysis and transaction data. *Clust Comput* 20(1):789–803
6. Nasukawa T, Yi J (2003) Sentiment analysis: capturing favorability using natural language processing. In: *Proceedings of the 2nd International Conference on Knowledge Capture*. ACM, pp 70–77
7. Appel O, Chiclana F, Carter J (2015) Main concepts, state of the art and future research questions in sentiment analysis. *Acta Polytech Hung* 12(3):87–108
8. Pang B, Lee L (2008) Opinion mining and sentiment analysis. *Found Trends Inf Retr* 2(1–2):1–135
9. Liu B (2012) Sentiment analysis and opinion mining. *Synth Lect Hum Lang Technol* 5(1):1–167
10. Pang B, Lee L, 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-Volume 10*. Association for Computational Linguistics, pp 79–86
11. Wilson T, Wiebe J, Hoffmann P (2005) Recognizing contextual polarity in phrase-level sentiment analysis. In: *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp 347–354

12. O'Hare N, Davy M, Bermingham A, Ferguson P, Sheridan P, Gurrin C, Smeaton AF (2009) Topic-dependent sentiment analysis of financial blogs. In: Proceedings of the 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion. ACM, pp 9–16
13. Go A, Bhayani R, Huang L (2009) Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, vol 1, no 12
14. Wu F, Yuan Z, Huang Y (2017) Collaboratively training sentiment classifiers for multiple domains. *IEEE Trans Knowl Data Eng* 29(7):1370–1383
15. Fernández AM, Esuli A, Sebastiani F (2016) Distributional correspondence indexing for cross-lingual and cross-domain sentiment classification. *J Artif Intell Res* 55(1):131–163
16. Wang L, Niu J, Song H, Atiquzzaman M (2018) SentiRelated: a cross-domain sentiment classification algorithm for short texts through sentiment related index. *J Netw Comput Appl* 101:111–119
17. Bader BW, Kegelmeyer WP, Chew PA (2011) Multilingual sentiment analysis using latent semantic indexing and machine learning. In: *IEEE 11th International Conference on Data Mining Workshops*, pp 45–52
18. Manek AS, Shenoy PD, Mohan MC, Venugopal KR (2017) Aspect term extraction for sentiment analysis in large movie reviews using Gini index feature selection method and SVM classifier. *World Wide Web* 20(2):135–154
19. Culnan M, McHugh P, Zubillaga J (2010) How large U.S. companies can use twitter and other social media to gain business value. *MIS Q Executive* 9(4):243–259
20. Di Gangi PM, Wasko M, Hooker RE (2010) Getting customers' ideas to work for you: learning from dell how to succeed with online user innovation communities. *MIS Q Executive* 9(4):163–178
21. He W, Zha S, Li L (2013) Social media competitive analysis and text mining: a case study in the pizza industry. *Int J Inf Manag* 33(3):464–472
22. Yang Y, Duan W, Cao Q (2013) The impact of social and conventional media on firm equity value: a sentiment analysis approach. *Decis Support Syst* 55(4):919–926
23. Phillips SJ, Anderson RP, Schapire RE (2006) Maximum entropy modeling of species geographic distributions. *Ecol Model* 190(3):231–259
24. Sun CJ, Yao L, Lin L, Sha XJ, Wang XL (2011) Semi-supervised biomedical relation classification using generalized expectation criteria. In: *2011 International Conference on Machine Learning and Cybernetics (ICMLC)*, vol 4, IEEE, pp 1949–1952
25. Mann GS, McCallum A (2010) Generalized expectation criteria for semi-supervised learning with weakly labeled data. *J Mach Learn Res* 11:955–984
26. Polat K, Güneş S (2009) A novel hybrid intelligent method based on C4. 5 decision tree classifier and one-against-all approach for multi-class classification problems. *Expert Syst Appl* 36(2):1587–1592
27. Schapire RE (2003) The boosting approach to machine learning: an overview. In: *Nonlinear estimation and classification*. Springer, New York, pp 149–171
28. Lewis DD (1998) Naive (Bayes) at forty: the independence assumption in information retrieval. In: *European Conference on Machine Learning*. Springer, Berlin, pp 4–15
29. Vapnik V (2013) The nature of statistical learning theory. Springer, Berlin
30. Levine R, Zervos S (1998) Stock markets, banks, and economic growth. *Am Econ Rev* 88:537–558