J Inf Process Syst, Vol.14, No.4, pp.904~915, August 2018
https://doi.org/10.3745/JIPS.04.0079

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

# Predicting the Unemployment Rate Using Social Media Analysis

Pum-Mo Ryu*

**Abstract**

We demonstrate how social media content can be used to predict the unemployment rate, a real-world indicator. We present a novel method for predicting the unemployment rate using social media analysis based on natural language processing and statistical modeling. The system collects social media contents including news articles, blogs, and tweets written in Korean, and then extracts data for modeling using part-of-speech tagging and sentiment analysis techniques. The autoregressive integrated moving average with exogenous variables (ARIMAX) and autoregressive with exogenous variables (ARX) models for unemployment rate prediction are fit using the analyzed data. The proposed method quantifies the social moods expressed in social media contents, whereas the existing methods simply present social tendencies. Our model derived a 27.9% improvement in error reduction compared to a Google Index-based model in the mean absolute percentage error metric.

**Keywords**

Google Index, Prediction, Sentiment Analysis, Social Media, Unemployment Rate

## 1. Introduction

Economic indicators released by official agencies are typically only available with a reporting lag of several weeks or a few months. It would be clearly helpful to have more timely forecasts of these indicators for sound economic decisions based on immediate big data statistics such as search engine queries or social media contents.

Many methods using search engine query data, such as Google Index (GI), have been proposed to forecast unemployment rates. Such methods have shown that there are strong correlations between search engine query data and unemployment rate and have demonstrated enhanced prediction results. In [1], a time-series causality approach using the well-known error correction model is employed to investigate the usefulness of the Google search activity data for predicting real economic behavior. The authors reported strong correlations between keyword searches and unemployment rate using monthly data from Germany.

In [2], the authors suggested the use of GI based on an Internet job search performed through Google as the best leading indicator to predict the US unemployment rate. Popular time series models


Manuscript received August 25, 2017; first revision November 15, 2017; accepted December 12, 2017.
**Corresponding Author:** Pum-Mo Ryu (pmryu@bufs.ac.kr)
* Dept. of ICT & Language Processing, School of Southeast Asian Studies, Busan University of Foreign Studies, Busan, Korea (pmryu@bufs.ac.kr)

augmented with GI have improved their out-of-sample forecasting performance. Pavlicek and Kristoufek [3] presented a data mining framework using search engine query data for unemployment rate prediction. Under the framework, a set of data mining tools including neural networks and support vector regressions are developed to forecast the unemployment trend.

Social media contents have been used as source for the analysis of social moods. By following the written expressions of individual users in social media over long periods of time, we are potentially able to infer details of sentiment dynamics such as happiness, anger, and grief. We hope to find the correlation between the analyzed social moods and the unemployment rate. In [4], the authors examined expressions made through tweets, uncovering and explaining temporal variations in the happiness and information levels over timescales ranging from hours to years. In [5], the quantified mood scoring and volume of keywords related to social media showed a significant relationship with the unemployment rate. For example, the volume of conversations categorized as showing confused moods in Ireland was correlated with the unemployment rate with lead time of three months, and conversations in the US regarding the loss of housing increased two months after unemployment spikes. In one study, [6] found that phrases containing terms such as "*home worse*," "*cough night*," "*sore head*," and "*swine flu*" closely tracked the reported ILI outbreaks statistics throughout the United Kingdom.

In the healthcare domain, scientists found that tweet streams have closely tracked reported cases of influenza-like illnesses, conditions that cause fever with cough or sore throat but which are not necessarily due to influenza, which has its own viral etiology. In [7], the authors found that terms including flu, swine, influenza, and infection, among many others, tracked user concerns during the H1N1 pandemic in 2009. They reported that tweet contents predicted flu outbreaks one to two weeks ahead of the surveillance average of the Centers for Disease Control and Prevention.

Most works on predictions using social media contents simply find leading indicators by measuring the correlation between outbreaks of keywords in social media and actual outbreaks of events. We suggest a novel method for generating numerical prediction values regarding the unemployment rate by analyzing social media contents. We suggest the use of social media as the leading indicator to predict the unemployment rate. We discover keywords whose frequency trends are closely correlated with the change in unemployment rate. We then build a prediction model using both the frequency trends of the keywords and data on the past unemployment rate. Both simple keyword frequency trends and sentiment-based keyword frequency trends are tested.

The contributions of this work for predicting the unemployment rate are as follows:

- The proposed model can reflect rapidly changing social phenomena as well as the current status of society, unlike a time-series prediction model that relies on past data. We hope to help social media correct the errors from the time gap of the source information. We also expect the time lag between the official reporting and unemployment rate to be shortened.
- Our method is challenging since the analyzed social media are quantified and evaluated by comparing actual unemployment rate data.

The rest of this paper is organized as follows: in Section 2, we describe the social media data used to predict the unemployment rate; in Section 3, we describe the natural language processing techniques to analyze social media contents; Section 4 discusses our prediction model; in Section 5, we describe the experiment and evaluation of the proposed model; finally, we present the conclusion in Section 6.

## 2. Data Collection

The data used in this paper come from three different sources. The monthly unemployment rate is released by Statistics Korea (http://www.kostat.go.kr). The exogenous variable in a previous study is the GI collected from Google Trends (http://www.google.com/trends). Finally, we explain the details of social media data. The social media used in our experiment include news articles, blogs, and tweets. News articles generated by media companies are one of the crucial sources of social opinions. Blogs usually written by the general public have increasingly become the mainstream for expressing public opinion, and they can provide information not expressed through the news. As an extremely popular online microblogging service, Twitter has a very large user base consisting of several millions of users. Each user submits periodic status updates known as tweets, which consist of short messages with maximum size of 140 characters. These updates typically consist of personal information about the users, news, or links to various contents such as images, videos, and articles. We can extract immediate insight into society based on the massive number of postings on users' lives.

For this study, news articles and web blogs were collected from ten major news media corporations and the Naver Blog site in Korea on a daily basis. More than 4,000 news articles and 58,000 blog postings were collected every day on average. We collected the tweets using a Twitter Streaming API that allows them to be searched based on keywords or strings in Korean. We collected 2.7M tweets per day on average. Fig. 1 shows the volume of our collection of news articles, blogs, and tweets on a monthly basis for 28 months (September 2011 through December 2013). The left y-axis is the volume of tweets, and the right y-axis is the volume of blogs and news. The volumes of blog and tweet collections have increased, but the volume of news collections is stable.

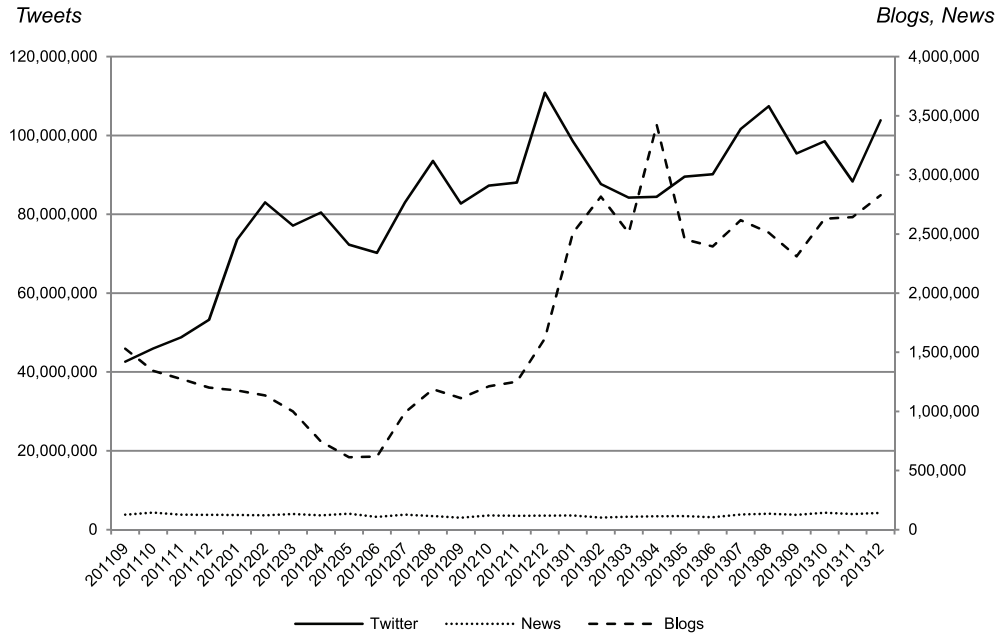


**Fig. 1.** Monthly frequency of news articles, blogs, and tweets written in Korean.

## 3. Natural Language Processing

Natural language processing (NLP) is an essential process for understanding social media content. To extract meaningful keywords and sentiments from such content, we performed part-of-speech tagging and sentiment analysis for Korean text.

Part-of-speech (POS) tagging is a process wherein a proper POS tag is assigned to each morpheme in a text. Because one morpheme can be mapped to multiple POS tags, the POS tagging system selects the best sequence of POS tags based on a probabilistic model or on lexico-morpheme rules. We exploit the Korean morphological analyzer in [8]. Simple tweet mentions and their POS-tagged results are presented in Table 1. The noun *주가(ju-ga)*(stock price) was selected as a keyword from Example 1. The noun *삼성(sam-sung)*(Samsung) and noun sequence *삼성주가(sam-sung ju-ga)* (stock price of Samsung) were selected as keywords from Example 2. The monthly frequency for the keywords was extracted from the POS-tagged social media content.

Sentiment analysis is one of the most popular research topics to which NLP and machine learning techniques can be applied. It is common to express this as a classification problem wherein a given text needs to be labeled as positive, negative, or neutral. Though classifiers built with machine learning algorithms have become commonplace in sentiment analysis, the core of many academic and commercial sentiment analysis systems is still the polarity lexicon, which can be constructed manually through heuristics or by using machine learning [9,10]. A lexicon-based method has advantages, i.e., we can easily grasp clues for the analysis results and tune the module by editing the lexicon. We built a Korean sentiment lexicon wherein positive and/or negative sentiments are assigned to the entries. Each entry of the lexicon is a text feature, either a morpheme or a combination of multiple morphemes; the existence of these features in a given text determines the sentiments of the text. To build the lexicon, we used a graph propagation algorithm based on the work in [9]. This algorithm propagates the sentiment polarity of the features on the graph built on large-sized social media content. We manually compiled a small seed sentiment lexicon because there is no widely accepted Korean sentiment lexicon such as the WordNet-Affect [11] or SentiWordNet [12] available in English. A lexicon-based sentiment analysis generally has high precision and low recall. In our forecasting task, high precision outweighs high recall because the prediction is sensitive to errors in the sentiment analysis, and the large volume of data may compensate for the low recall problem. Example 1 presents a sentence with *짜증나(jja-jeung-na)* (annoying), a negative polarity expression. Example 2 presents a sentence with "^^" (happy), a positive expression (see Table 1).

Most of the news articles present facts for particular entities or events; hence the difficulty of extracting sentiment expressions from news articles. Instead, people express their opinions including beliefs, emotions, speculations, and so on in blogs and tweets. Therefore, we mainly tackled blogs and tweets as important sources for a sentiment analysis.

We focused on the emotional details of keywords for our prediction model using a sentiment analysis technique. Modeling the co-occurrence of positive, negative, or neutral expressions of a keyword gives another view of the keyword. For example, the keyword *주가(ju-ga)*(stock price) co-occurs with the negative expression *짜증나(jja-jeung-na)*(annoying) in Example 1. In this case, *주가(ju-ga)* is a keyword causing negative moods for society, whereas *주가(ju-ga)* co-occurs with the positive expression "^^" (happy) in Example 2, with *주가(ju-ga)* creating a positive mood for society.

**Table 1.** Example sentences for POS tagging and sentiment analysis

| | |
|---|---|
| Example 1 | |
| Tweet mention (Korean) | *주가가* (*ju-ga-ga*) *떨어져서* (*tteol eo jyeo seo*) *짜증나* (*jja jeung na*) |
| English translation | It's annoying because the stock price is falling. |
| POS-tagged | *주가*/nc+*가*/jc *떨어지*/pv+*어서*/ec *짜증나*/pv+*아*/ef |
| Sentiment analysis | Polarity: NEGATIVE<br>Clue expression: *짜증나* (*jja jeung na*) (annoying) |
| Example 2 | |
| Tweet mention (Korean) | *삼성* (*sam-sung*) *주가* (*juga*) *오르네* (*oreu-ne*) ^^ |
| English translation | Samsung's stock price goes up ^^ |
| POS-tagged | *삼성*/nc *주가*/nc *오르*/pv+*네*/ef ^^/s |
| Sentiment analysis | Polarity: POSITIVE<br>Clue expression: ^^/s |

# 4. Prediction Method

In this section, we define the data model used in prediction models, illustrate the basic prediction models, describe a CCF analysis to select data that are highly correlated with the unemployment rate, and finally show the result of fitting the selected data with the prediction models.

## 4.1 Data Model

The data models used in this paper are as follows:

- *Unemployment Rate Index* (*UI*): The monthly unemployment rate index of Korea. We denote the unemployment rate in $t^{th}$ month as $y_t$. $UI=\{y_t | t=1,2,3,\ldots,T\}$.
- *Google Index* (*GI*): The monthly Google search keyword frequency of a keyword provided by Google Trends. We denote the Google Index of keyword $w$ in $t^{th}$ month as $g_{w,t}$. $GI(w)=\{g_{w,t} | t=1,2,3,\ldots,N\}$.
- *Social Keyword Index* (*SKI*): The monthly social media frequency of a keyword extracted from social media content. We denote the Social Keyword Index of keyword $w$ in $t$-th month in media m as $f_{w,m,t}$. $SKI(w)=\{f_{w,m,t} \mid t = 1,2,3,\ldots,T;\ m = \text{news}|\text{blogs}|\text{tweets}\}$.
- *Social Sentiment Index* (*SSI*): The monthly social media sentiment frequency of a keyword extracted from social media content. We denote the Social Sentiment Index of keyword $w$ in $t^{th}$ month in media $m$ of certain sentiment $s$ as $f_{w,m,s,t}$. $SSI(w)=\{f_{w,m,s,t} \mid t=1,2,3,\ldots,T;\ m = \text{news}|\text{blogs}|\text{tweets};\ s = \text{pos}|\text{neg}\}$.

## 4.2 Prediction Model

As one of the most useful methodologies for analyzing a time series, the autoregressive integrated moving average (ARIMA) model offers great flexibility in analyzing various time series and gives accurate forecasts [13]. The ARIMA model with seasonal terms (SARIMA) can be written as follows:

$$\emptyset_p(B)\Phi_P(B^s)(1-B^s)^D(1-B)^d y_t = \delta + \theta_q(B)\Theta_Q(B^s)e_t \quad (1)$$

where $\phi_P(B)$, $\theta_q(B)$, $\Phi_P(B)$, $\Theta_Q(B)$ are as follows:

$$\emptyset_p(B) = 1 - \emptyset_1 B^1 - \emptyset_2 B^2 - \cdots - \emptyset_p B^p$$
$$\theta_q(B) = 1 - \theta_1 B^1 - \theta_2 B^2 - \cdots - \theta_q B^q$$
$$\Phi_P(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps}$$
$$\Theta_Q(B^s) = 1 - \Theta_1 B^s - \Theta_2 B^{2s} - \cdots - \Theta_Q B^{Qs}$$

where $B$ is the backshift (lag) operator (e.g., $B^b Z_t = Z_{t-b}$) and $e_t$ is white noise $WN(0, \sigma_e^2)$. For example, the ARIMA($p$,$q$) model can be simply described as follows:

$$y_t = \delta + \emptyset_1 y_{t-1} + \cdots + \emptyset_p y_{t-p} - \theta_1 e_{t-1} - \cdots - \theta_q e_{t-q} + e_t \quad (2)$$

where $y_t$ is the unemployment rate of $t^{th}$ month and et is the error terms in $t^{th}$ month. Consequently, the AR($p$) model is simply expressed as follows:

$$y_t = \delta + \emptyset_1 y_{t-1} + \cdots + \emptyset_p y_{t-p} \quad (3)$$

As an extended version of the ARIMA model, the ARIMAX model also includes other independent (predictor) variables. The ARIMAX model is similar to a multivariate regression model but allows taking advantage of the autocorrelation that may be present in the residuals of the regression to improve the accuracy of a forecast. An ARIMAX model simply adds the covariate on the right-hand side of ARIMA as follows:

$$\emptyset_p(B)\Phi_P(B^s)(1-B^s)^D(1-B)^d y_t = \delta + \theta_q(B)\Theta_Q(B^s)e_t + \beta x_{t-d} \quad (4)$$

where $x_t$ is a covariate at time $t$ and is its coefficient. For brevity, we use only a single covariate in the model above, but more than two covariates can be contained in the model as an additive type. The ARIMAX model can be considered a special case of the transfer function model. Just like the ARIMA model, the ARIMAX model without a seasonal factor—including more than $k$ (≥2) covariates—can be expressed as follows:

$$y_t = \delta + \beta_1 x_{1,t-d_1} + \cdots + \beta_k x_{k,t-d_k} + \emptyset_1 y_{t-1} + \cdots + \emptyset_p y_{t-p} - \theta_1 e_{t-1} - \cdots - \theta_q e_{t-q} + e_t \quad (5)$$

One of the advantages of the ARIMAX model over ARIMA is that it uses the information of the covariate series. Practically, however, the choice of lag d is not easy, especially when more than two covariates are contained in the model. A simple and useful model incorporating the historical information and covariate information is (seasonal) autoregressive with exogenous variables (ARX) model. The ARX model is a linear difference equation model that relates the input to the output. By increasing the number of exogenous input terms, we can better approximate the observed dynamics in the systems. The ARX model is defined as follows:

$$y_t = \delta + \beta_1 x_{1,t-d_1} + \cdots + \beta_k x_{k,t-d_k} + \emptyset_1 y_{t-1} + \cdots + \emptyset_p y_{t-p} + e_t \quad (6)$$

where $y_{t-1}$ ($i$=1, …, 12) is the historical time series of lag $i$, $x_{i,t-di}$ ($i$=1,…,p) is the covariate time series, and $e_t$ is an error term. Eq. (6) without historical variables can be regarded as a regression-type model. All coefficients of Eq. (6) can be estimated from the ordinary least squares method.

## 4.3 CCF Analysis

Next, we describe the procedure used to select keywords whose trends are highly correlated with the unemployment rate using a cross-correlation function (CCF). We collected a set of keywords related to the unemployment rate from 10 persons who submitted 100 keywords each. The set consists of 622 keywords. We extracted the GI, SKI, and SSI of each keyword and compared the indices with the UI using CCF in the R package (http://www.r-project.org/).

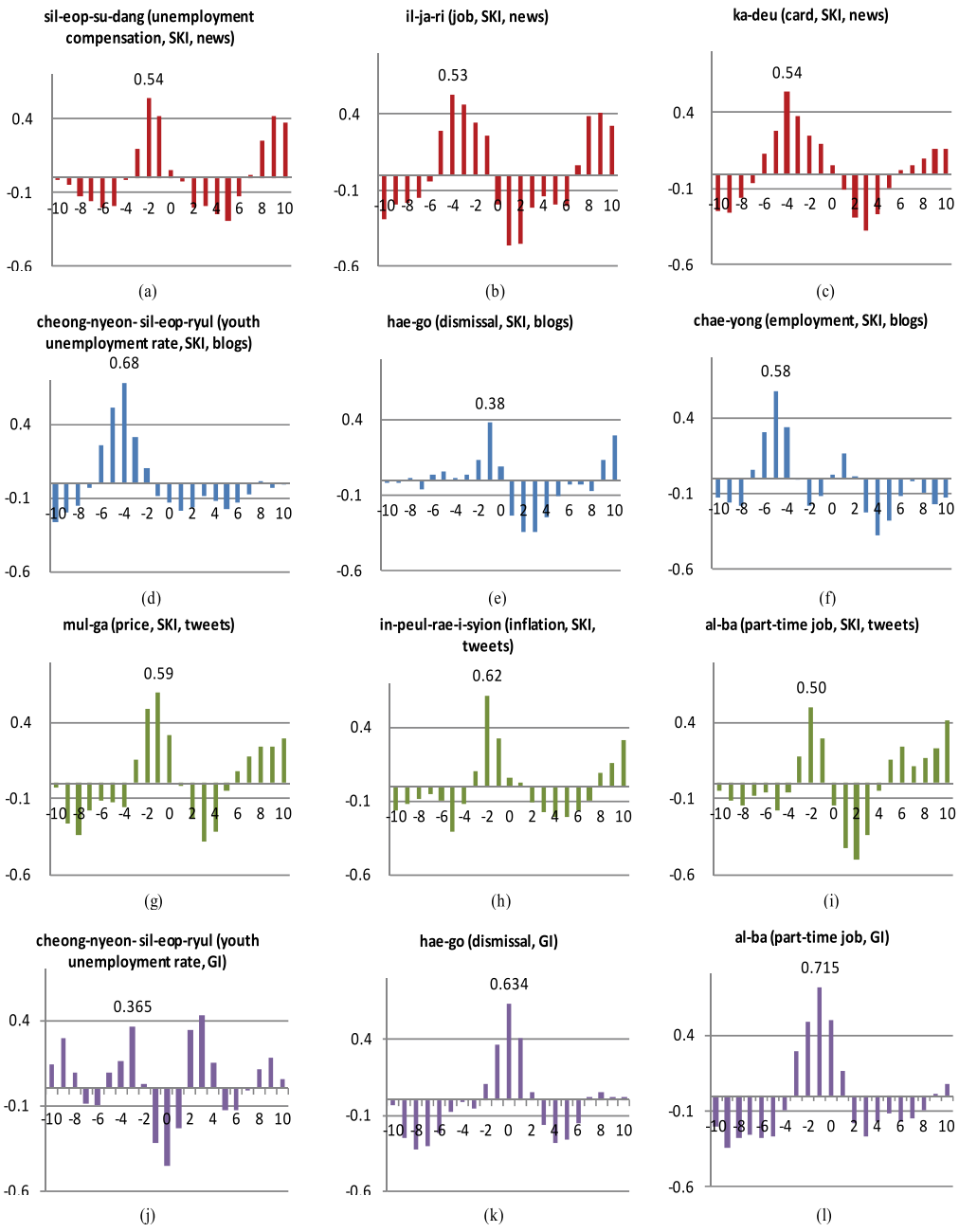


**Fig. 2.** CCF analysis results of keywords. Their SKI and GI are compared with the UI.

The problem we are considering is the description and modeling of the relationship between two time series, such as (GI(*w*), UI) pair, (SKI(*w*), UI) pair, or (SSI(*w*), UI) pair. In the relationship between two time series ($y_t$ and $x_t$), CCF identifies the lags of the x-variable that might be useful predictors of $y_t$. CCF is defined as a set of sample correlations between $x_{t+h}$ and $y_t$ for h = 0, ±1, ±2, ±3, and so on. A negative value for *h* is a correlation between the x-variable at time before *t* and the y-variable at time *t*. For instance, $h = -2$ means that the CCF value gives the correlation between $x_{t-2}$ and $y_t$. The goal is to identify $x_t$ whose keyword is leading and whose keyword is lagging. The range of correlation is inclusive between +1 and −1, where +1 is a total positive correlation, 0 is no correlation, and −1 is a total negative correlation. Fig. 2 shows parts of the CCF analysis results of the keywords. Their SKI and GI are compared with the UI. The first nine graphs (*a* through *i*) show the CCF results for the SKI for several keywords and the UI. The last three graphs (*j* through *l*) show the CCF results for the GI and UI. For example, the SKI for *인플레이션*(*in-peul-rae-i-syion*)(inflation) showed the highest correlation with the UI at −2 time interval in the tweet collection (*h*). The unemployment rate will likely increase after 2 months if the frequency of *인플레이션*(*in-peul-rae-i-syion*) increases in the tweets. As another example, the GI for *해고*(*hae-go*)(dismissal) showed the highest correlation with the UI at 0 time interval. This means that the frequency of *해고*(*hae-go*) as a search query for Google is tightly coupled with the unemployment rate with time lag of 0.

We selected the keywords whose CCF shows high correlation with the UI and whose time lag is between 0 and −4. The selected keywords were used as covariates in the prediction models.

## 4.4 Model Fitting and Prediction

We fitted the ARIMA, ARIMAX, and ARX models using the UI, GI, and SI, respectively, using the data selected by the CCF analysis. We tested all possible combinations of keywords and selected the best models in each model category as follows:

- *Model_U*: The ARIMA model (Eq. (3)) based on the UI, excluding GI, SKI, and SSI as exogenous variables. The fitted model is shown in Eq. (7):

$$\hat{y}_t = 1.1497 + 0.6398 \cdot y_{t-1} \tag{7}$$

- *Model_G*: The ARIMAX model (Eq. (5)) based on the UI including GI as an exogenous variable. The UI of the previous month and GI for *청년실업률*(*cheong-nyeon-sil-eop-ryul*)(youth unemployment rate) and *해고*(*hae-go*)(dismissal) are used to fit the model. If we know the GI for *청년실업률*(*cheong-nyeon-sil-eop-ryul*) 3 months earlier and *해고*(*hae-go*) 1 month earlier, we can predict the unemployment rate for this month using Eq. (8).

$$\hat{y}_t = 1.1240 - 0.5783 \cdot y_{t-1} - 0.0016 \cdot g_{cheong-nyeon-sil-eop-ryul,t-3} + 0.0043 \cdot g_{hae-go,t-1} \tag{8}$$

- *Model_K*: The ARX model (Eq. (6)) based on the UI including SKI as an exogenous variable. We fit three models based on different media types: news, blogs, and tweets. Eq. (9) shows a prediction model fitted by the SKI of tweets. The frequency of *물가*(*mul-ga*) (price) and *인플레이션*(*in-peul-rae-i-syion*) (inflation) in the tweets is used.

$$\hat{y}_t = 1.1631 + 0.4020 \cdot y_{t-1} + 0.2172 \cdot f_{mul-ga,tweets,t-1} + 4.1153 \cdot f_{in-peul-rae-i-syion,tweets,t-2} \quad (9)$$

- *Model_S*: The ARX model (Eq. (6)) based on the UI including SSI as an exogenous variable. This model consists of three models based on the media types: news, blogs, and tweets. Eq. (10) shows a fitted model where three SSIs for three keywords— 실직(*sil-jik*) (unemployment), 알바(*al-ba*) (part-time job), and 주가(*ju-ga*) (stock price)—are used. The lag and sentiment for the three keywords are all different from each other.

$$\hat{y}_t = 1.6823 + 0.3116 \cdot y_{t-1} - 82.6107 \cdot f_{sil-jik,tweets,pos,t-3} + 1.4283 \cdot f_{al-ba,tweets,neg,t-2} + 4.5683 \cdot f_{ju-ga,tweets,neg,t-1} \quad (10)$$

*Model_U* and *Model_G* are the baselines in our experiment. As our proposed models, *Model_K* and *Model_S* test the usability of social media content in predicting the unemployment rate.

## 5. Experimental Results

We built the prediction models using data for 2 years from September 2011 to October 2013 and tested the models using data for four months from September 2013 to December 2013. The data have a sufficient time span to model the seasonal characteristics of the unemployment rate. *Model_U*, *Model_G*, and three types of *Model_K* and *Model_S* were compared based on their goodness-of-fit (GOF) and prediction accuracy. *Model_U* and *Model_G* are the baselines in this experiment. GOF was measured using data for two years, with the prediction accuracy measured based on data for four months. We also compared our models with the forecast of Trading Economics (http://www.tradingeconomics.com), a commercial economic data provider that provides predictions for the unemployment rate for each country on a monthly basis.

The GOF and prediction accuracy of models were evaluated using the well-known prediction metrics of mean absolute error (MAE), root mean square error (RMSE), and mean absolute percentage error (MAPE) as in Eqs. (11), (12), and (13), respectively. Let $y_i$ and $f_i$ be the actual data and corresponding predicted data, respectively, with n denoting the number of predictions made. The metrics derive a lower value (≥0) when the given model shows good performance. MAE gives an average absolute value of the difference between the estimated forecast and the actual value. RMSE assigns a relatively high weight to large errors, since the errors are squared before they are averaged. RMSE is most useful when large errors are particularly undesirable. MAPE is a more objective statistical indicator because the measurement is in relative percentage and will not be affected by the unit of the forecasting series.

$$\mathrm{MAE}(M) = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| \quad (11)$$

$$\mathrm{RMSE}(M) = \sqrt{\frac{\sum_{i=1}^{n}(f_i - y_i)^2}{n}} \quad (12)$$

$$\mathrm{MAPE}(M) = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{f_i - y_i}{y_i}\right| \quad (13)$$

In view of GOF (Table 2), the MAPE of the proposed models is between 2.801 and 5.627. This result indicates that we can trust the predicted results with more than 94% confidence level. *Model_U* showed the highest error rate among all models for all metrics. This means that the exogenous variables GI and SI improved the performance of the other models including *Model_G*, *Model_K*, and *Model_S*. In particular, social-media based models *Model_K* and *Model_S* showed a lower error rate than *Model_G*. We can infer that social media contents are more effective in grasping social moods on the topic of labor than GI. *Model_S* showed the lowest RMSE and derived the most stable results without showing sharp error pitches. It reduced the errors by 48.5% compared to *Model_G*. *Model_K* (blogs) showed the lowest MAE and MAPE among all models. Compared to *Model_G, Model_K* (blogs) reduced the errors by 42.0% and 41.0% in the MAE and MAPE metrics, respectively.

**Table 2.** Evaluation results: goodness of fit of all models and improvement over *Model_G*

| Model | RMSE | Improvement in RMSE (%) | MAE | Improvement in MAE (%) | MAPE | Improvement in MAPE (%) |
|---|---|---|---|---|---|---|
| *Model_U* (baseline 2) | 0.279 | - | 0.192 | - | 5.629 | - |
| *Model_G* (baseline 3) | 0.222 | - | 0.158 | - | 4.748 | - |
| *Model_K* (news) | 0.165 | 25.6 | 0.138 | 12.6 | 4.230 | 10.9 |
| *Model_K* (blogs) | 0.137 | 38.0 | 0.092 | 42.0 | 2.801 | 41.0 |
| *Model_K* (tweets) | 0.201 | 9.2 | 0.154 | 2.3 | 4.810 | -1.3 |
| *Model_S* | 0.114 | 48.5 | 0.099 | 37.2 | 3.182 | 33.0 |

**Table 3.** Evaluation results: prediction accuracy of all models and improvement over *Model_G*

| Model | RMSE | Improvement in RMSE (%) | MAE | Improvement in MAE (%) | MAPE | Improvement in MAPE (%) |
|---|---|---|---|---|---|---|
| Trading economics (baseline 1) | 0.415 | - | 0.375 | - | 13.664 | - |
| *Model_U* (baseline 2) | 0.317 | - | 0.298 | - | 10.832 | - |
| *Model_G* (baseline 3) | 0.433 | - | 0.403 | - | 14.586 | - |
| *Model_K* (news) | 0.358 | 17.3 | 0.342 | 15.2 | 12.349 | 15.3 |
| *Model_K* (blogs) | 0.448 | -3.5 | 0.390 | 3.3 | 14.245 | 2.3 |
| *Model_K* (tweets) | 0.318 | 26.5 | 0.291 | 27.8 | 10.513 | 27.9 |
| *Model_S* | 0.946 | -118.6 | 0.879 | -118.0 | 31.603 | -116.7 |

In view of the prediction accuracy (Table 3), the three models of *Model_K* showed better performance than *Model_G* for the MAE and MAPE metrics. In particular, *Model_K* (tweets) showed the lowest MAE and MAPE among all models. It reduced the errors by 27.8% and 27.9% for the MAE and MAPE metrics, respectively, compared to *Model_G*. The frequency of the keywords extracted from tweets can be assumed to be related closely to the unemployment rate index. Note, however, that *Model_S* showed a higher error rate than other forecasts, unlike the case of GOF. Therefore, the coverage and precision of the sentiment analysis should be increased when we apply the technique to a real-world application. When the emotional analysis result is applied to the prediction of unemployment rate, the reason for the relatively low performance is the low performance of emotional analysis. There is also the problem

of objective analysis being difficult because of the short forecast period. Therefore, future research is needed to improve the accuracy of sentiment analysis. *Model_K* (news) and *Model_K* (tweets) showed better performances than the Trading Economics forecast. As such, our models can be applied to a commercial service. Social media including news and tweets showed better performance than that of blogs in prediction because news and tweets reflect actual social phenomena.

Fig. 3 shows the actual unemployment rate and the fitted and prediction results of the other model. The graphs of the first 24 months are the fitted results, and the graphs of the last four months are the predicted results. The overall fitted graphs show reasonable results, but some of the predicted graphs show irregular patterns. Note, however, that the predicted data by *Model_K* (news) and *Model_K* (tweets) show the same patterns (down-up-down-up) as the UI data for the prediction period. Blogs are good for model fitting but show a low accuracy in terms of prediction compared to news or tweets as in Table 3. We should track the prediction accuracy for a larger number of months to find out the characteristics of social media types in our work.

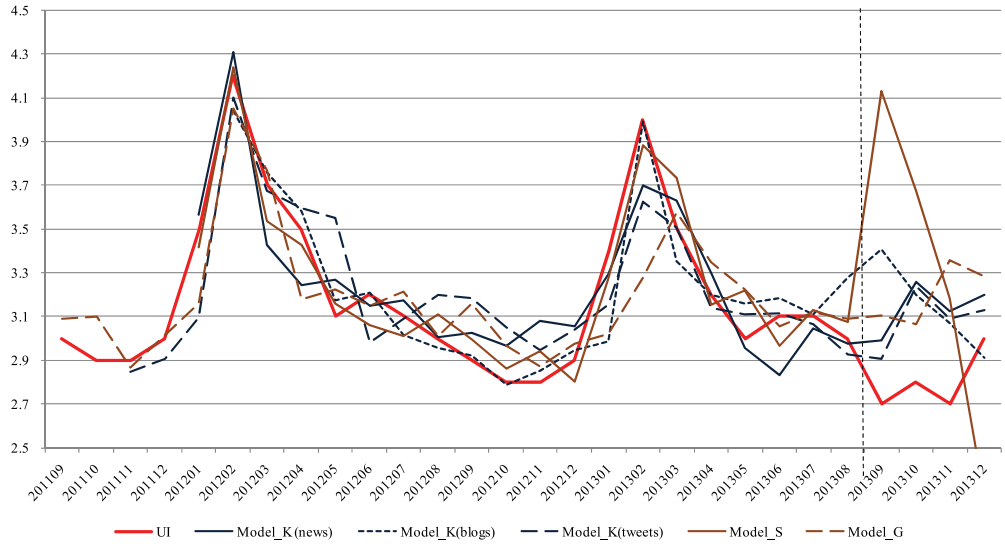


**Fig. 3.** Comparison of actual unemployment rate (UI) and fitted/predicted data by models. The graphs for the first 24 months are for fitted data, and the graphs for the last four months are for predicted data. The y-coordinate represents the monthly unemployment rate.

## 6. Conclusion

We presented several models to predict the unemployment rate based on social media analysis. We showed the effectiveness of social media in analyzing and quantifying social moods by applying the data to the unemployment prediction models. Our models derived better results than the GI-based model and simple time-series model. We will apply social media analysis to predict other social indices such as consumer price index or consumer sentiment index. Because such indices are tightly coupled with public life, a large number of mentions regarding these indices will be posted in social media. To apply such analysis, the sentiment analysis for Korean should be improved.

## Acknowledgement

This work was supported by the research grant of the Busan University of Foreign Studies in 2018.

## References

[1] N. Askitas and K. F. Zimmermann, "Google econometrics and unemployment forecasting," *Applied Economics Quarterly*, vol. 55, no. 2, pp. 107-120, 2009.

[2] F. D'Amuri and J. Marcucci, "'Google it!' Forecasting the US unemployment rate with a Google job search index," *FEEM Working Paper No. 31*, 2010.

[3] J. Pavlicek and L. Kristoufek, "Nowcasting unemployment rates with google searches: evidence from the visegrad group countries," *PloS One*, vol. 10, no. 5, article no. e0127084, 2015.

[4] P. S. Dodds, K. D. Harris, I. M. Kloumann, C. A. Bliss, and C. M. Danforth, "Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter," *PloS One*, vol. 6, no. 12, article no. e26752, 2011.

[5] United Nations Global Pulse, "Using social media to add depth to unemployment statistics," UN Global Pulse White Paper, 2011.

[6] V. Lampos and N. Cristianini, "Nowcasting events from the social web with statistical learning," *ACM Transactions on Intelligent Systems and Technology*, vol. 3, no. 4, article no. 72, 2012.

[7] A. Signorini, A. M. Segre, and P. M. Polgreen, "The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic," *PloS One*, vol. 6, no. 5, article no. e19467, 2011.

[8] S. Lim, C. Lee, P. M. Ryu, H. Kim, S. K. Park, and D. Ra, "Domain-adaptation technique for semantic role labeling with structural learning," *ETRI Journal*, vol. 36, no. 3, pp. 429-438, 2014.

[9] L. Velikovich, S. Blair-Goldensohn, K. Hannan, and R. McDonald, "The viability of web-derived polarity lexicons," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Stroudsburg, PA: Association for Computational Linguistics, 2010, pp. 777-785.

[10] K. J. Lee, J. E. Kim, and B. H. Yun, "Extracting multiword sentiment expressions by using a domain-specific corpus and a seed lexicon," *ETRI Journal*, vol. 35, no. 5, pp. 838-848, 2013.

[11] C. Strapparava and A. Valitutti, "WordNet affect: an affective extension of WordNet," in *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, Lisbon, Portugal, 2004, pp. 1083-1086.

[12] A. Esuli and F. Sebastiani, "SentiWordNet: a publicly available lexical resource for opinion mining," in *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy, 2006, pp. 417-422.

[13] G. E. P. Box and G. M. Jenkins, *Time Series Analysis: Forecasting and Control*. Englewood Cliffs, NJ: Prentice Hall, 1976.

**Pum-Mo Ryu** https://orcid.org/0000-0002-9777-9211

He received the B.S. degree in computer engineering from Kyungpook National University, Daegu, South Korea in 1995 and the M.S. degree in computer engineering from POSTECH, Pohang, Korea, in 1997. He received the Ph.D. degree in computer science from KAIST, Daejeon, Korea, in 2009. Currently he is an associate professor in Department of ICT & Language Processing, School of Southeast Asian Studies, Busan University of Foreign Studies, Busan, Korea. His research interests include natural language processing, text mining, knowledge engineering and question answering.