ARIMAX and ARX Models with Social Media Information to Predict Unemployment Rate

Kaaen Kwon
Department of Business Data Convergence, Chungbuk National University, Cheongju, Korea
Email: ggaaen@gmail.com

Wan-Sup Cho[1] and Jonghwa Na[2]
[1] Department of Management Information System, Chungbuk National University, Cheongju, Korea
[2] Department of Information & Statistics/Business Data Convergence, Chungbuk National University, Cheongju, Korea
Email: wscho63@gmail.com, cherin@chungbuk.ac.kr

***Abstract*—To keep current with trends in the society, social media has been actively used for understanding issues and moods of real world. In this paper, we suggest the method using the social media to predict the unemployment rate based on natural language processing and statistical modelling. We adopted AutoRegressive Integrated Moving Average with eXogenous variables (ARIMAX) and AutoRegressive with eXogenous variables (ARX) model to predict the unemployment rate and compared our model to a Google Index based model. Our model derived 27.8% and 27.9% improvements in error reduction compared to an existing model in mean absolute error and mean absolute percentage error metrics, respectively.**

***Index Terms*—social media, unemployment rate, prediction, sentiment analysis, Google Index**

I. Introduction

Economic indicators such as consumer price index, stock market prices, retail sales and unemployment rate are only available with a reporting lag of several weeks or a few months. In the rapidly changing world situation, those would be one step behind in the world of trends to predict unemployment rate. To overcome the limit, social media information has been widely discussed on the usage as predicting economic variables in recent. Social media information has a powerful advantage that it is easy to gain and be spread rapidly in a real time and is often connected to real-world event.

Some of works on predicting unemployment rate based on search engine query data have been proposed that there are strong correlations between Google Index (GI) and unemployment rate [1]-[3]. Social media contents also contain sentimental expression of individual users. Therefore, we can find a simple correlation between the trend and unemployment rate [4], [5]. However, social media contents have been hardly used as predicting variables to predict the unemployment rate. In the paper, we use keywords extracted from social media as the leading keywords and then build a prediction model using both the frequency trends of the keywords and data on the past unemployment rate. We tested both simple keywords frequency trends and sentiment-based keyword frequency trends.

The remainder of this paper is organized as follows:

Related works on predicting unemployment rate are surveyed in the next section, then the system overview in the third section. In the fourth section the proposed methods including description of the social media data used to predict the unemployment rate and model fitting. In the fifth section, we perform the experiment and analyze the results. Finally, the conclusion is provided.

II. Related Works

Predicting unemployment rate correctly is very important because it helps government policy respond flexibly to the labour market and the nation's economy. Therefore, many works on predicting unemployment rate has been discussed and the methods using search engine query data have been proposed in recent. One of the works in Germany used unemployment and economy search data of GI and has discovered that adopting GI variables to Error Correction Model (ECM) would make the model improve [1]. And another works demonstrated that generally ARX and ARMAX with GI variables as an eXogenous variables showed the smallest MSE among about 500 models including AR and ARMA models to predict the US unemployment rate [2].

III. Predicting Unemployment Rate System Overview

The predicting unemployment rate system is comprised of the following two frameworks: 1) Social Media Processing Framework, and 2) Prediction Framework, as shown in Fig. 1.

Manuscript received December 20, 2014; revised May 22, 2015.

doi: 10.12720/joams.4.5.401-404

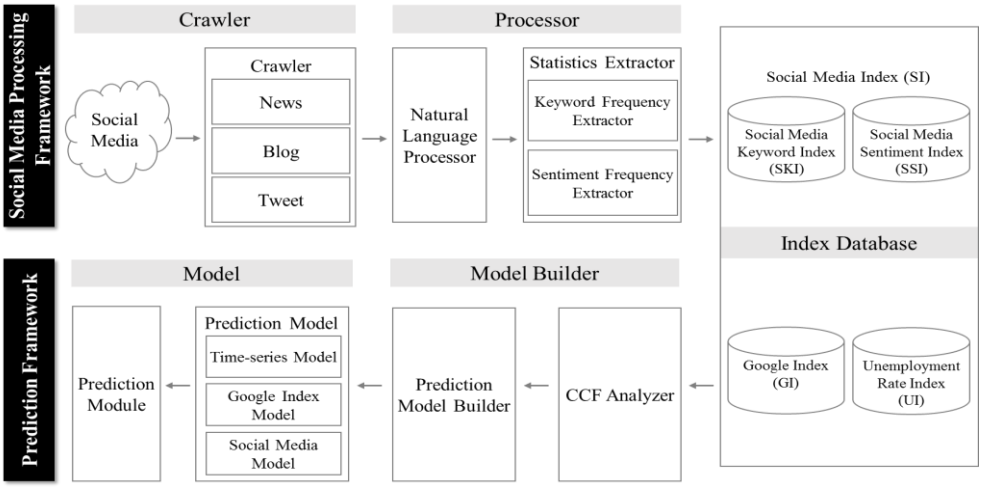

Figure 1. Prediction framework and social media processing framework

A. *Social Media Processing Framewor*

The social media processing framework collects news articles, blogs, and tweets and generates an index database for the prediction framework using natural language processor and statistics extractor modules. The index database is composed of a social media index (SI), Google index (GI) and unemployment rate index (UI). The statistics extractor extracts keyword frequencies and builds a social media keyword index (SKI) on a monthly basis. The negative/positive sentiment frequency of the keywords is also extracted and is saved in the social media sentiment index (SSI). The GI for candidate keywords is retrieved from the Google Trends site [6]. The unemployment rate index (UI) of Korea is retrieved from the Statistics Korea site [7].

B. *Prediction Framework*

The prediction framework consists of a cross correlation function (CCF) analyzer, a prediction model builder, and a prediction module. Highly correlated keywords are selected by CCF analyzer and the model builder builds three types of prediction model such as time-series model, GI model and social media model.

IV. PROPOSED MODEL

A. *Data*

For this study, news articles and Web blogs are collected from ten major news media corporations and the Naver blog site [8] in Korea on a daily basis for 28 months (Sep. 2011 through Dec. 2013). Tweets are collected using a Twitter Streaming API on a same period and only tweets written in Korean are concerned. More than 4,000 news articles, 58,000 blog postings and 2.7M tweets are collected every day on average. We also used unemployment rate which is published by Statistics Korea every month.

B. *Prediction Method*

1) *ARIMAX model*

Between Autoregressive Integrated Moving Average (ARIMA) and ARIMAX model, there is one difference which is an ARIMAX model simply adds the covariate, as follows:

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

where $\emptyset(B), \theta_c(B), \Phi(\mathrm{B}), \Theta_c(B)$ are as follows:

$$\emptyset(B) = 1 - \emptyset_1 B - \emptyset_2 B^2 - \cdots - \emptyset_p B^p$$

$$\theta_q(B) = 1 - \theta_q B - \theta_2 B^2 - \cdots - \theta_q B^q$$

$$\Phi(\mathrm{B}) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps}$$

$$\Theta_q(B) = 1 - \Theta_1 B^s - \Theta_2 B^{2s} - \cdots - \Theta_Q B^{Qs}$$

where B is backshift(lag) operator (e.g., $B^b Z_t = \mathrm{Z}_{t-b}$),$\mathrm{e}_t$is white noise $\mathrm{WN}(0, \sigma_e^2)$ and$x_t$is a covariate at time t and is its coefficient. For brevity, we use only a single covariate in the above model, but more than two covariates can be contained in the model as an additive type.

2) *ARX model*

The autoregressive with the covariate (ARX) model is a simple and useful model incorporating the historical information and covariate information defined as follows:

$$\begin{aligned}y_t = \delta + \beta_1 x_{1,t-d_1} + \cdots + \beta_k x_{k,t-d_k} \\ +\emptyset_1 y_{t-1} + \cdots + \emptyset_p y_{t-p} \\ -\theta_1 e_{t-1} - \cdots - \theta_q e_{t-q} + e_t\end{aligned} \quad (2)$$

where $y_{t-i}(i = 1, \ldots, 12)$ is the historical time series of lag i, $x_{i,t-d_i}$ (i =1,…,p) is the covariate time series, and et is an error term. Equation (2) without historical variables can be regarded as a regression type model. All coefficients of equation (2) can be estimated from the ordinary least squares (OLS) method.

C. *Corss Correlation Function Analysis*

The keywords whose trends are markedly correlated to the unemployment rate using a cross correlation function (CCF) are collected from ten persons who submitted 100 keywords each. We extracted the GI, SKI, and SSI of each keyword, and compared the indexes to the UI using CCF in the R package [9]. The keywords whose CCF shows a high correlation with the UI and whose time lag

is between 0 and -4 are selected. If the time lag is positive, the keywords are post indicators for the unemployment index. If the absolute time lag of a keyword is too large (>4), its usability is low. The selected keywords are used as covariates in the prediction models.

D. Model Fitting and Prediction

First, we used CCF analysis for selecting the data and then we fitted the ARIMAX, and ARX models using the UI, GI, and SI, respectively. All possible combinations of keywords are tested and we select the best models in each model category as follows:

Model_G: The ARIMAX model (equation1) based on the UI including the GI as an exogenous variable. In equation (8), the UI of the previous month and the 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) of three months before, and 해고 (hae-go) of one month before, we can predict the unemployment rate of this month using the generated equation.

$$\begin{aligned}\hat{y}_t &= 1.1240 - 0.5783 \times y_{t-1} \\ &\quad -0.0016 \times g_{cheongnyeon\ sileopryul,t-3} \\ &\quad +0.0043 \times g_{haego,t}\end{aligned} \tag{3}$$

Model_K: The ARX model (equation2) based on the UI including the SKI as an exogenous variable. We fit three models based on different media types: news, blogs, and tweets. Equation (4) 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.

$$\begin{aligned}\hat{y}_t &= 1.1631 + 0.4020 \times y_{t-1} \\ &\quad -0.2172 \times f_{ulga,tweets,t-1}\end{aligned} \tag{4}$$

$$+4.1153 \times f_{\in peulraeisyion,tweet,t-2}$$

Model_S: The ARX model (equation2) based on the UI including the SSI as an exogenous variable. This model comprises three models based on the media types: news, blogs, and tweets. Equation (5) 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 to each other.

$$\begin{aligned}\hat{y}_t &= 1.68231 + 0.31162 \times y_{t-1} \\ &\quad -82.61068 \times f_{siljik,tweets,pos,t-3} \\ &\quad +1.42832 \times f_{alba,tweets,neg,t-2} \\ &\quad +4.56829 \times f_{juga,tweets,neg,t-1}\end{aligned} \tag{5}$$

Model_G is the model introduced in [2] and is the baseline in our experiment. Model_K and Model_S are our proposed models. We want to show the usability of social media content in predicting the unemployment rate compared to the Google Index or the original unemployment rate index.

V. Exprimental Results

For experiment, we split the data. One is for building the prediction models with two years data, Sep. 2011 to Oct. 2013, and another is for testing the models with four months data, Sep. 2013 to Dec. 2013. Model_G, and three types of Model_ K and Model_S were compared based on their goodness-of-fit (GOF) and prediction accuracy. Model_G isthe baseline in this experiment.

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).

TABLE I. Goodness of Fit of All Models and Improvement Over Model_G

Model	RMSE	Imprv. (RMSE)	MAE	Imprv. (MAE)	MAPE	Imprv. (MAPE)
Model_G (baseline)	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%

GOF is shown in Table I and Model_G showed the highest error rate of all models for all metrics. We can infer that social media contents are more effective in catching social moods on the topic of labour than Google Index.

Prediction accuracy is shown in Table II and the proposed models showed meaningful results. Model_K shows a better accuracy than Model_G for most metircs. However, Model_S showed a much higher error rate than other forecasts, unlike the case of GOF.We can infer that the analysed data do notrepresentthe current social moods in terms coverage and accuracy. The coverage and precision of the sentiment analysis should therefore be increased when we apply the technique to a real-world application.If we improve the performance of a sentiment analysis technique to the degree of POS tagging for Korean, we can achieve a much higher or more competitive accuracy than Model_K.

VI. Conclusions

In the paper, we proposed models to predict unemployment rate using social media information. We fitted ARIMAX model and ARX model using the UI, GI, and SI, respectively and showed the improvement of model accuracy compared to models that used GI. Also, we designed predicting unemployment rate system from collecting social data to building predicting models. It can help to better predict unemployment rate and the system process would apply to another things such as analysis society issues, an economic indicator and so on.

In the future works, other social variables, such as the consumer price index or consumer sentiment index will be added as eXogenous variables. Making the predicting unemployment rate system automatic and deriving results in a real time will be future works as well.

TABLE II. Prediction Accuracy of All Models and Improvement Over Model_G

Model	RMSE	Imprv. (RMSE)	MAE	Imprv. (MAE)	MAPE	Imprv. (MAPE)
Model_G (baseline)	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%

Acknowledgment

This research was supported by the MSIP(The Ministry of Science,ICT and Future Planning), Korea, under the "SW master's course of a hiring contract" support program (NIPA-2014-HB301-14-1011)

References

[1] N. Askitasand and K. F. Zimmermann, "Google econometrics and unemployment forecasting," in *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," *MPRA Paper*, no. 18248, pp. 1-52, 2009.
[3] Xu. Wei, Z. Li, C. Cheng, and T. T. Zheng, "Data mining for unemployment rate prediction using search engine query data," *Service Oriented Computing and Applications 7*, no. 1, pp. 33-42, 2013.
[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, 2011.
[5] UNGP, "Using social media to add depth to unemployment statistics," *UN Global Pulse White paper*, 2011.
[6] Google Trends. [Online]. Available: http://www.google.com/trends
[7] Statistics Korea. [Online]. Available: http://www.kostat.go.kr
[8] Naver. [Online]. Available: http://www.naver.com
[9] The R Project for Statistical Computing. [Online]. Available: http://www.r-project.org/

Kaaen Kwon received her B.S. degrees at the Department of Information Mathematics and Industrial Engineering from Korea University, Korea in 2013. She is in Master degree at the Department of Business Data Convergence from Chungbuk National University in Korea. His research interest include bigdata, data mining and social data analysis.

Jonghwa Na received his M.S. and Ph.D. degrees at the Department of Statistics from Seoul National University, Korea. He is a professor at the Department of Management Information & Statistics/Business Data Convergence, Chungbuk National University in Korea. His research interests include big data, data mining, statistical modeling/prediction and social data analysis.

Wan-Sup Cho received his M.S. and Ph.D. degrees at the Department of Computer Science from KAIST, Korea, in 1987 and 1996, respectively. He is a professor at the Department of Management Information System, Chungbuk National University in Korea. His research interests include database, bigdata, business intelligence and ERP.