Prediction of Increase in COVID-19 Confirmed Cases Through Multilingual Keyword Extraction

Gi-Hu Kim and Gil-Jin Jang*
School of Electronic and Electrical Engineering, Kyungpook National University, Daegu 41566, Korea
climate1022@gmail.com, gjang@knu.ac.kr
**correspondence author*

Abstract

This paper proposes new method for COVID-19 spread prediction using the multi-lingual news. A novel keyword extraction method is proposed and a random forest classifier that takes the relative frequency of keyword appearances as input. According to the experimental results, the average test AUC (area under the ROC curve) scores of United States, Republic of Korea, and Japan showed 7.67% improved performance on the average compared to the existing system.

Keywords: Keyword extraction, COVID19, document classification.

1. Introduction

COVID-19 is a disease that spreads rapidly because it is high infective contagious. The government of the Republic of Korea adjusted the alert level for COVID-19 from severe to vigilant on May 11, 2023, but the risk of this highly infectious disease has not disappeared in Korea. In order for governments to respond promptly, it is very important to detect when the number of confirmed cases increases rapidly [1]. As COVID-19 was a pandemic with worldwide risks, there are many studies to accurately predict it [2,3]. Various prediction methods are have been studied, such as a predictive model based on real-time data of confirmed cases worldwide [2] or applying machine learning techniques [3]. However, since these real-time statistics-based methods require the actual number of confirmed cases as input, there is a limitation that it is impossible to predict areas where aggregated information is inaccurate or impossible to collect due to the characteristics of the highly contagious and rapidly spreading coronavirus.

Recently, a social network service (SNS) such as Twitter, filters news articles related to coronavirus using natural language processing technology, and a SNS Big Data analysis model is presented that predicts the outbreak of confirmed cases using information from these news for 7 days [4]. A deep learning model that predicts the onset of COVID-19 using Seq2Seq Attention and Word2Vec keyword time sires data has also been proposed [5]. Most recently, a deep learning model to predict the outbreak of COVID19 by automatically indexing and classifying multilingual articles related to COVID-19 using ChatGPT and multilingual BERT has also been proposed [6]. However, even if the onset of COVID-19 is predicted, due to the nature of COVID-19, which is highly contagious and mutates quickly, handling new terms describing new symptoms and characteristics required. It is not easy to cope with new words only by classifying and classifying them, and it takes a lot of cost and time to build a new system to prepare for them.

In this paper, we propose a system that automatically extracts keywords from multilingual news and predicts the increase in the number of confirmed cases by country based on the extracted words. This system proposes an extraction method that improves the performance of the existing proposed key word extraction method by country. In addition, based on news data for each country, key words related to the number of confirmed cases are automatically extracted. A binary classifier for predicting the increase in the number of confirmed cases by country was constructed using random forest with the probability information of each country's key words. The proposed key word extraction method can predict the increase or decrease in the number of confirmed cases for a total of 3 countries in Japanese, Korean, English, and can achieve an average AUC (area under the ROC curve) of 11% or more for each country compared to the existing key word extractor.

The structure of this thesis is as follows. Section II describes the proposed system and method in detail. Section III shows the experimental results of automatic extraction of key words in multilingual articles, and concludes in Section IV.

2. Proposed Method

A. Definition of Keydate

In this paper, we do not directly predict the number of confirmed cases, but predict the point at which the increase in the number of confirmed cases exceeds the standard set in the paper. The criterion for judging the increase in confirmed cases is based on the fact that the incubation period for COVID19 is 2 weeks, and if the number of confirmed cases on that day increases by 10% or more compared to the average number of confirmed cases in the previous 2 weeks, it is judged to be the keydate. Using Johns Hopkins University's real-time cumulative number of confirmed cases by country [7], the average number of confirmed cases for 2 weeks is compared with the number of confirmed cases on the day, and if it is below the standard value, it is classified as 0, and if it is above the standard, it is classified as 1. Therefore, the prediction of the increase in confirmed cases can be defined as a binary pattern classification problem in which each day is a key increase or not, and the correct answer of the pattern classifier is used.

B. Proposed keyword extraction-based increasement detection model Training and Inference

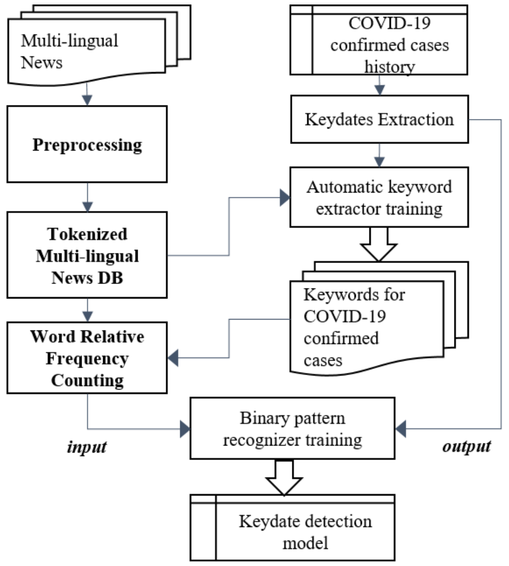
Multi-lingual News
Preprocessing
Tokenized Multi-lingual News DB
Word Relative Frequency Counting
COVID-19 confirmed cases history
Keydates Extraction
Automatic keyword extractor training
Keywords for COVID-19 confirmed cases
Binary pattern recognizer training
input
output
Keydate detection model

Fig. 1 Training confirmed cases increment detection model using the proposed automatic relative-keyword extraction method.

In this study, we propose a model for detecting the increase in confirmed cases based on the relative frequency of occurrence of key words in each country's news. Figure 1 shows the learning process of the proposed detection model. First, the keydate information is classified by labeling it as 0 and 1 from the information on the number of confirmed cases. News articles for each country are tokenized into words in noun units using spaCy [8], one of the natural language processing libraries, and Regular Expression for each country language. The keywords related to the increase in the number of confirmed cases are extracted by matching keydate labeled '1' with vocabs extracted from news by country for that day. A set of news articles combined with noun-level words is defined as vocab, and the number of vocabs varies depending on the number of news articles per country. Therefore, in order to make it independent of the number of news articles, it is calculated using the relative frequency of appearance, which indicates how many keywords appear in one news article, and uses this as an input to the binary pattern classifier. The goal of the binary pattern classifier is to learn to determine whether a given day is a keydate or not based on the relative frequency of keywords entered as input.

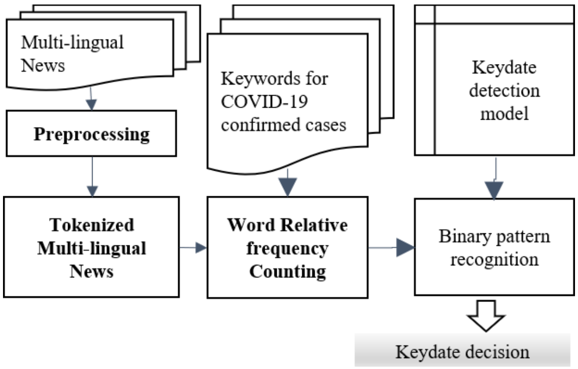
Multi-lingual News
Preprocessing
Tokenized Multi-lingual News
Keywords for COVID-19 confirmed cases
Word Relative frequency Counting
Keydate detection model
Binary pattern recognition
Keydate decision

Fig. 2 Inferring confirmed cases increment based on the trained keydate detection model.

Figure 2 shows the process of judging whether or not it is a key increase from the news article of the date composed of letters using the binary pattern recognizer learned in Figure 1 above. From vocab, a set of keywords obtained through learning in Figure 1, and key words of news articles of the corresponding date, the input of the binary pattern classifier is created using the method of calculating the relative probability of occurrence of words. The input obtained in this way is used in the detection model to infer whether or not it is keydate.

C. Previous Keyword Extraction Method

In [9], A method for extracting key words from news articles was proposed as follows. A set of documents collected based on keydate is defined as $D^{(kd)}$, and the number of appearances of the word t_i for this set is defined as TC (Term count), and the formula is as follows.

$$TC_i^{(kd)} = \sum_{d_j \in D^{(kd)}} n_{k,j} \tag{1}$$

The importance of the word t_i can be defined by normalizing $TC_i^{(kd)}$, which means the number of occurrences of a word, with a relative frequency of occurrence. This is defined as $NTFKD_i$, and the formula is as follows [9]:

$$NTFKD_i = tanh\left(\alpha \frac{TC_i^{(kd)}}{\sum_j^{|D|}\sum_k^{|T_j|} n_{k,j}} \right) \quad , \qquad (2)$$

$$\alpha = 30000$$

In Equation (2), the denominator is the number of occurrences of all words in all documents, and serves to normalize $TC_i^{(kd)}$, which is the number of occurrences of the word t_i, to a relative frequency. A positive constant α appropriately reduces the value of a relatively very large denominator, and 30,000 was used in [9]. In addition, the function tanh is used to approximate $NTFKD_i$ as a conditional probability against the key increase date, and it is normalized so that it is not dependent on a specific word with a very high frequency. However, when this method is used for multilingual news articles, the corpus size $\sum_j^{|D|}\sum_k^{|T_j|} n_{k,j}$ used as the denominator affects the matrix of $NTFKD_i$, and thus affects the performance of the binary classification model, it is necessary to readjust α.

D. Proposed Keyword Extraction Method

Since α of $NTFKD_i$ needs to be adjusted for each news article for each country, we propose $RTCKD_i$ (Relative Term Count in KeyDate) that is not dependent on hyperparameter α. $RTCKD_i$ which formulates the relative frequency of occurrence of keywords by country, is as shown in Equation (3) below.

$$RTCKD_i = \left(\frac{TC_i^{(kd)}}{\sum_j^{|D|}\sum_k^{|T_j|} n_{k,j}} \right) \approx P(t_i|D^{(kd)}) \qquad (3)$$

$RTCKD_i$, the relative frequency of occurrence of keywords in news by country, means the ratio of occurrences of word t_i to the overall corpus size of the news. Unlike $NTFKD_i$ used in [9], this frequency of occurrence does not depend on the hyperparameter α, so it does not need to be readjusted each time a binary classification model is trained for each country.

3. Experimental Result

A Google News Search crawl was used to search and collect articles related to COVID19. Considering the epidemic period of COVID19, the total cumulative number of days was 1308, limiting the range from August 1, 2019 to February 28, 2023. Train/test was divided at a ratio of about 8:2 from the total number of cumulative days, and the actual period and number of days are shown in Table 1.

Table 1. Training and Test data split

	Duration	Days
Train	2019/08/01 - 2022/06/30	1065
Test	2022/07/01 - 2023/02/28	243
Total	2019/08/01 - 2023/02/28	1308

In order to accurately predict the increase in confirmed cases, an optimal set of keywords used for prediction is required. It can be seen that the optimal set of keywords differs according to the key word extraction method, and the number of optimal keyword sets is different for each country. As for the key word extraction method, the NTFKD presented in [9] and the RTCKD presented in this paper were applied, and the number of key words was increased by 5 from 5 to 30, extracting key words as training data and parameters of the random forest classifier used in [9]. was imported and trained in the same way, and evaluated using the 5-fold cross validation method. As an evaluation scale, AUC (Area Under the Curve), one of the commonly used metrics for binary classifiers, was used [10]. Table 2 shows the AUC performance of NTFKD and RTCKD by country. The optimal number of keyword sets showing the best performance in NTFKD was 20, 25, and 10 in the order of English, Japanese, and Korean, and 25, 25, and 5 in RTCKD. Also, the average performance of multilingual AUC was improved by 7.67% compared to the existing proposed method. It can be seen that the optimal set of keywords differs according to the key word extraction method, and the number of optimal keyword sets is different for each country.

Table 2. AUC performances on test data with the optimal number of keyword for NTFKD, RTCKD

Test AUC	United States	Japan	Korea
NTFKD	0.79(C=20)	0.73(C=25)	0.54(C=10)
RTCKD	**0.81(C=25)**	**0.79(C=25)**	**0.69(C=5)**

Table 3 is the result of comparing the performance of each method with the optimal key words of English actually extracted. NTFKD's optimal number of keywords is 20, which is 5 fewer than RTCKD's, but RTCKD's AUC performance is 2% higher than NTFKD's. More specialized keywords were extracted from NTFKD, but it is more effective that we were able to more accurately predict the increase and decrease in COVID-19 confirmed cases even among general keywords.

Table 3. Extracted Keywords and AUC performances on test data with the optimal number of keywords for NTFKD, RTCKD

	NTFKD	RTCKD
C	20	25

Extracted Keywords	monkeypox distancing covid delta coronavirus lockdown pcr quarantine pandemic swab mrna abortion leukemia psoriasis cohort droplet influenza receptor sinus airway	people time patient covid day symptom health disease study treatment body life week child risk death month infection family pain doctor condition blood virus person
Test AUC	0.79	0.81

4. Conclusion

In this study, we proposed an automatic key word extraction method based on multilingual news and a method for predicting the increase in the number of confirmed cases based on a binary classifier. It can be seen that the optimal set of keywords differs according to the key word extraction method, and the number of optimal keyword sets is different for each country. While the existing proposed method [9] is effective only for English, the method proposed in this paper is effective not only for English, but also for Korean and Japanese, and shows improved performance in English as well. Therefore, it will be applied to predicting the increase in the number of confirmed cases in real time rather than the existing method, enabling rapid response to COVID-19.

Acknowledgments

This research was supported by the Government-wide R&D Fund for Infections Disease Research (GFID), funded by the Ministry of the Interior and Safety, Republic of Korea (grant number: 20016180).

References

[1] J.-S. Kim, “Infectious Disease: Past, Present, and Future,” *Vacuum Magazine*, 7(2), pp. 13-17, 2020.

[2] Huang, J., Zhang, L., Liu, X., Wei, Y., Liu, C., Lian, X., Huang, Z., Chou, J., Liu, X., Li, X., & others (2020). Global prediction system for COVID-19 pandemic. Science bulletin, 65(22), 1884.

[3] Malki, Z.; Atlam, E.-S.; Ewis, A.; Dagnew, G.; Ghoneim, O.A.; Mohamed, A.A.; Abdel-Daim, M.M.; Gad, I. The COVID-19 pandemic: Prediction study based on machine learning models. *Environ. Sci. Pollut. Res.* 2021, *28*, 40496–40506

[4] Azzaoui, A. E., Singh, S. K., & Park, J. H. (2021). SNS big data analysis framework for COVID-19 outbreak prediction in smart healthy city. *Sustainable Cities and Society*, *71*, 102993.

[5] Kim, Y., Park, C.R., Ahn, J.P., & Jang, B. (2023). COVID-19 outbreak prediction using Seq2Seq+ Attention and Word2Vec keyword time series data. Plos one, 18(4), e0284298.

[6] Seungtae Kang and Gil-Jin Jang. " COVID-19 Multilingual News Article Auto-indexing and Classification using ChatGPT and Multilingual BERT " Journal of the Institute of Electronics and Information Engineers 2023, vol.60, no.7, pp. 20-29

[7] “New COVID-19 cases worldwide,” https://coronavirus.jhu.edu/data/new-cases, accessed on 25 May 2023

[8] M. Honnibal, and I. Montani, *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*, 2017.

[9] H. Jeong, G.-J. Jang. “Automatic Extraction of Keywords for COVID-19 Cases Increment Detection” Journal of the Institute of Electronics and Information Engineers, to be published

[10] Jeong Kyun Kim, Kang Bok Lee, Sang Gi Hong. "ECG-based Biometric Authentication Using Random Forest" Journal of the Institute of Electronics and Information Engineers 2017, vol.54, no.6, pp. 100-105.