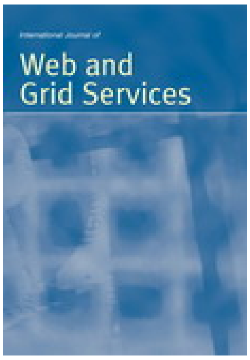


International Journal of Web and Grid Services

ISSN online: 1741-1114 - ISSN print: 1741-1106
https://www.inderscience.com/ijwgs

Sentiment analysis and counselling for COVID-19 pandemic based on social media

Ha-Young Lee, Ok-Ran Jeong

DOI: 10.1504/IJWGS.2023.10054493

Article History:

Received:	04 May 2022
Last revised:	31 August 2022
Accepted:	01 September 2022
Published online:	06 March 2023

Sentiment analysis and counselling for COVID-19 pandemic based on social media

Ha-Young Lee and Ok-Ran Jeong*

School of Computing,
Gachon University,
Seongnam-si, Gyeonggi-do, 13120, Korea
Email: hhzet11@gachon.ac.kr
Email: orjeong@gachon.ac.kr
*Corresponding author

Abstract: As COVID-19 emerged and prolonged, various changes have occurred in our lives. For example, as restrictions on daily life are lengthening, the number of people complaining of depression is increasing. In this paper, we conduct a sentiment analysis by modelling public emotions and issues through social media. Text data written on Twitter is collected by dividing it into the early and late stages of COVID-19, and emotional analysis is performed to reclassify it into positive and negative tweets. Therefore, subject modelling is performed with a total of four datasets to review the results and evaluate the modelling results. Furthermore, topic modelling results are visualised using dimensional reduction, and public opinions on COVID-19 are intuitively confirmed by generating representative words consisting of each topic in the word cloud. Additionally, we implement a COVID-chatbot that provides a question-and-answer service on COVID-19 and verifies the performance in our experiments.

Keywords: social media analysis; sentiment analysis; topic modelling; COVID-chatbot; Google BERT; Microsoft DialoGPT.

Reference to this paper should be made as follows: Lee, H-Y. and Jeong, O-R. (2023) ‘Sentiment analysis and counselling for COVID-19 pandemic based on social media’, *Int. J. Web and Grid Services*, Vol. 19, No. 1, pp.34–57.

Biographical notes: Ha-Young Lee is currently an MS student in the School of Comping at Gachon University, South Korea. She is a member of the Intelligent Data Analysis Laboratory (IDALab) in the School of Computing at Gachon University. She is interested in natural language processing, social media mining, machine learning, and deep learning.

Ok-Ran Jeong is currently a Professor in the School of Computing, Gachon University, South Korea. Her current research interests include big data mining, machine learning, deep learning, and applications of artificial intelligence. She received her PhD in Computer Science and Engineering from the Ewha Womans University in 2005. She was a postdoctoral researcher at the University of Illinois at Urbana-Champaign, USA and Seoul National University, Korea. She joined the faculty of the Department of Software Design and Management at Gachon University in 2009.

1 Introduction

In January 2020, the World Health Organization declared an international public health emergency due to the coronavirus, and a pandemic was underway worldwide. As a result, numerous events were postponed and cancelled, and as of March 2022, more than 400 million confirmed cases and more than 6 million deaths occurred. So we would like to find out how people think and feel about COVID-19, which has such a huge social and economic impact on the world.

With the recent rapid growth of social media, numerous users produce large amounts of data in real-time through social media based on smart devices. So, it is possible to collect, analyse, and process big data generated based on social media. Accordingly, research using analysis technologies that analyse users' opinions and thoughts based on data collected from social media, extract meaningful information or predict the flow of a specific field is being actively conducted in various fields (Yun et al., 2017; Manguri et al., 2020).

This study aims to collect tweets, which are posts related to COVID-19, through Twitter and analyse people's emotions. Data collected from social media has a high proportion of subjective opinions, so it has good conditions to perform sentiment analysis. The number of studies that conduct sentiment analysis using social media data has increased, and based on the research analysis results, it is being used in various fields such as prediction, recommendation, and chatbot (Kim et al., 2020; Nhung et al., 2016; Tyagi and Tyripathi, 2019; Yoo et al., 2020).

We tried to analyse people's thoughts and feelings about COVID-19 by classifying Twitter data collected through sentiment analysis into positive and negative tweets and then conducting topic modelling for each emotion. Based on the results of this study, if applied to a chatbot service that can comfort and heal people's emotions and produces answers suitable for the context, it will be very useful in solving many questions and depression caused by COVID-19.

Our main contribution is summarised as follows:

- We conduct sentiment analysis through the collected Twitter data. It is possible to find out what emotions dominate the dataset in each of the early and late periods of COVID-19. In addition, an emotional score is imposed to find out whether people are positive or negative about COVID-19 overall. It also checks how the distribution of emotions changes in the early and late stages.
- For each dataset classified as period and emotion polarities, topic modelling is used to examine what topic is predominant. The modelling results are visualised through dimension reduction, and the representative words constituting each topic are generated in the word cloud to intuitively check the public thoughts.
- As COVID-19 continues in the long term, we implement the COVID-chatbot to solve unresolved questions and depression problems. We generate topic and emotion labels and train them together with special tokens in a pre-trained language model to generate more effective responses than existing models.

The remainder of this paper is organised as follows. In Section 2, we introduce related work. In Section 3, we describe the datasets used in the study and the pre-processing process conducted for experiments. In Sections 4 and 5, we explain the sentiment analysis and the topic analysis. In Section 6, we present the implementation of the COVID-chatbot and the results of several experiments. Finally, we summarise our conclusion and provide some ideas for future work.

2 Related work

2.1 Text mining

Text mining is a technique based on natural language processing (NLP) technology and is a research technique that extracts and analyses patterns or relationships from unstructured text data to derive valuable information. Representatively, discussions on coronavirus issues have recently been active on social media such as Twitter and Facebook, and many studies have been conducted. 40,000 tweets including 20 corona-related hashtags were collected to analyse representative topics through LDA machine learning techniques (Xue et al., 2020), and tweets related to corona vaccines were collected to analyse different topics and emotional trends by country (Yousefinaghani et al., 2021). Furthermore, posts related to COVID-19 vaccines were collected through Facebook posts, and emotional trends and discussion topics were predicted through deep learning technology (Hussain et al., 2021).

Since the posted text of data produced by social media such as Twitter is text data, text mining can be used for social media data analysis. Major technologies include morpheme analysis and word frequency analysis (Park et al., 2018).

- *Morphological analysis:* Morpheus is the smallest unit of speech with a certain meaning, and morpheme analysis extracts only meaningful morpheme from the corpus, which identifies linguistic attributes from words. There are representative techniques such as lemmatisation to restore the original form of words and stemming to extract the word's stem (Sun et al., 2014).

  In this paper, we go through the process of returning the word to the basic type through lemmatisation, and for example, if 'am', 'is', or 'are' are present, we convert it to 'be'. We also use stemming to remove the endings from the words and extract the word's stem to solve the problem of counting different words if there are multiple uses of words with the same meaning.

- *Word frequency analysis:* term frequency (TF) is a process of identifying parts of words through morpheme analysis, and then extracting parts of words with meanings such as nouns, verbs, and adjectives to determine how many each word appeared in the document. Through this process, we can look at the flow of full-text data, and, the higher the frequency of a word's appearance, the more the word corresponds to the keyword. Currently, since the stopwords generally show a high frequency of appearance in sentences, a process of removing stopwords is essential to successfully proceed with the word frequency analysis. Frequency analysis also involves visualising analysis results through word clouds that emphasise the term by expressing words that appear relatively frequently in certain documents relative to other words (Kim et al., 2017).

  After analysing social media data related to COVID-19, visualisation of modelling results is performed through the word cloud.

2.2 Topic analysis

- *Topic modelling:* topic modelling is one of the text mining techniques, a method of automatically extracting topics of words contained in the extracted document. The most used algorithm is latent Dirichlet allocation (LDA), an unsupervised learning algorithm that finds hidden topics in documents and binds them together by document and keyword. That is, assuming the distribution of Dirichlet, it represents the distribution in which each document and word is assigned to the topic by utilising the latent variable called the topic (Kim et al., 2017).

  Unlike the other clustering techniques, where one document is assigned to only one topic, topic modelling can be said to be a more suitable technique for modelling in the real world because one document can correspond to multiple topics simultaneously (Heo and Yang, 2020). In conclusion, topic modelling learns a topic vector represented by words, which is as high-dimensional as the number of words, and visualises it to express the results of topic modelling.

  We would like to find out about people's perception and dominant topics of COVID-19 through topic modelling and how people's perception of the early and late periods has changed.

- *Topic modelling evaluation criteria:* there are two main evaluation criteria for the LDA model. First, perplexity refers to the degree of confusion and represents a value for how well a specific probability model predicts the observed value. This score means that the smaller the value, the better the modelling results reflect the document (Williams et al., 2020).

  In contrast, coherence measures the consistency of a topic, and the better the modelling, the more similar words are gathered within a topic. By calculating the similarity between words from this value, we can see how many semantically similar words are gathered in the subject. However, if the value of the coherence result is too high, the amount of information decreases, causing the topic becomes monotonous. However, on the contrary, if the coherence is too low and there is no association between information, the meaning of the analysis becomes low.

  There are also several methods for calculating coherence scores, and we utilise C_v and C_umass techniques. First, C_v is evaluated based on sliding windows, one set of parent words, indirect confirmation measurements using normalised point-wise mutual information (NPMI), and cosine similarity, with values between 0 and 1. This scale has proven particularly appropriate for evaluating the quality of a subject based on other widely used topic consistency measures and large empirical comparisons and provides the score closest to human evaluation. On the other hand, C_umass defines scores based on the number of simultaneous occurrences of documents, first-order divisions, and log conditional probabilities as a confirmation scale, with values between –14 and 14. This value means that the closer it is to zero, the more complete consistency it has (Zoya et al., 2021).

2.3 Pre-trained language model

- *BERT:* BERT is a NLP pre-training technology developed by Google, and is a language model that performs well in all NLP fields, not limited to specific fields. It is an abbreviation for bidirectional encoder representations from transformer, a language model that can improve the performance of certain models through pre-trained embeddings. This model is based on the transformer, but unlike other models, it uses only transformer encoders, introducing self-attention. Thus, it consists of the sum of three input embeddings: token, segment, and position embedding, and uses the masked language model (MLM) and next sentence prediction (NSP) to ensure that the properties of the language are well learned. As a result, state-of-the-art (SOTA) was achieved in 11 NLP tasks such as classification and QA (Devlin et al., 2018).

  We intend to classify the topics and emotions of the conversation dataset using BERT along with the LDA of topic modelling. Thus, the COVID-chatbot that we implement is trained to generate answers suitable for the context of emotions and conversations so that it can solve many questions and depression caused by COVID-19.

- *DialoGPT:* DialoGPT is an abbreviation for the dialogue generative pre-trained transformer (GPT), an extension of the GPT-2 developed by Microsoft. As with GPT-2, DialoGPT is an autoregressive language model, and since it uses a transformer of multi-layer as its model configuration, so it exhibits excellent performance in sentence generation (Radford et al., 2019). Unlike GPT-2, however, DialoGPT is a model that aims to generate human-like natural text in conversation by learning through large-scale dialogue pairs extracted from the Reddit discovery chain. This language model also achieves SOTA in the fields of automatic and human evaluation and can generate answers with performance like human-generated responses (Zhang et al., 2019; Mehri and Eskenazi, 2020).

  We intend to implement a chatbot using the DialoGPT language model to generate consistent and plentiful answers when people ask several questions related to Corona or complain of depression.

3 Tweet dataset

People generate large amounts of data in real-time through social media. By collecting and processing the generated big data, meaningful information such as public opinions may be extracted. We use social media data that share individual subjective and honest opinions to find out about the public's feelings and perceptions of COVID-19.

3.1 Dataset

Twitter data among social media data are used to find out people's thoughts about COVID-19 (Banda et al., 2021). The dataset contains tweet data related to COVID-19 starting March 22, 2020, consisting of the ID of the tweet and the date and time it was created. Since tweets cannot be distributed due to Twitter's policy, the dataset is constructed by crawling the tweets based on the tweet ID through a hydrator.

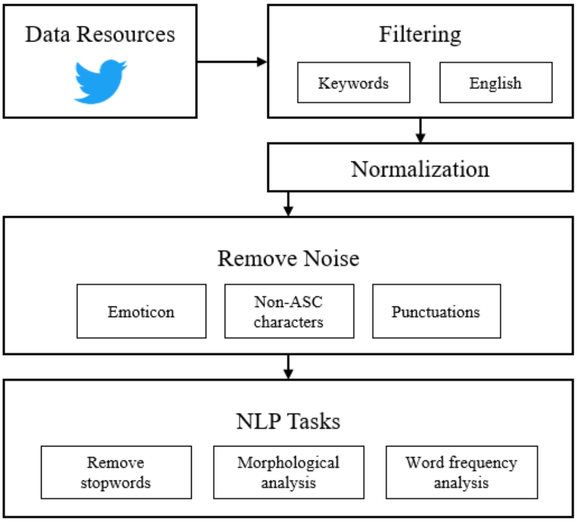
To find out how people's perception of COVID-19 has changed as COVID-19 is prolonged, the initial period of COVID-19 (2020.03.22 – 2020.04.04, 14 days) and the later period (2021.09.22 – 2021.10.0, 14 days) are set, and tweet data for that period are collected. At this time, all tweets related to COVID-19 written in the period are crawled through a dataset containing the id of the tweet from which the re-tweet was removed, and if the tweet publisher deleted the tweet, it is reflected. As a result, 12,378,260 tweets were finally collected during the early period of COVID-19, and 2,514,487 tweets were collected during the late period of COVID-19.

3.2 Pre-processing

A pre-processing process is performed for text mining analysis of tweet data (Lyu et al., 2021). First, only tweet data written in English is filtered, and tweets including those keywords are filtered by designating 'covid19', 'COVID-19', 'corona', 'coronavirus', 'virus', 'pandemic', 'outbreak', and 'Wuhan' as keywords. Subsequently, all data were changed to lowercase letters, and the process of removing emoticons was performed. In addition, URLs and references, numbers, and symbols that become noise in NLP are removed, and all non-English-classified data among tweets written in English and other languages are removed. It also uses the 'string.functions' library to remove the punctuation symbol.

Following the pre-processing process for full-scale NLP, use the stopwords in the 'nltk' library to remove the stopwords and then, remove all previously set keywords. At this time, the sentence symbols and stopwords that are not well removed are additionally designated and removed together. Subsequently, tweets written in only one letter through the pre-processing process do not play a large role, so all of them are removed. Finally, we proceed with lemmatisation to restore the original form of the word and stemming to extract the word's stem.

Figure 1 Structural diagram of the pre-processing process (see online version for colours)



As a result, the number of tweets that can be used in the early period of COVID-19 was 2,862,842, and 689,915 were used in the later period, and the results can be seen in Table 1.

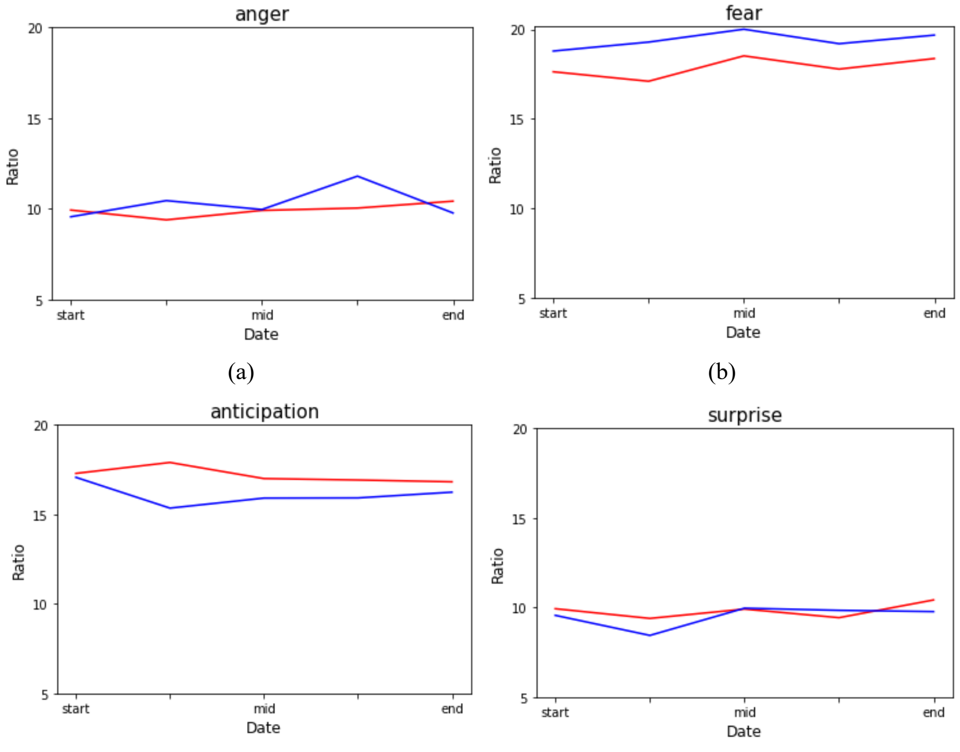
Table 1 Number of tweets after collection and pre-processing

Task	*Early period*	*Late period*
Origin	12,378,260	2,514,487
After filtering	5,282,142	1,286,176
After NLP task	2,862,842	689,915

4 Sentiment analysis

We utilise sentiment analysis to extract meaningful information from collected and pre-processed datasets. During each period, we check what emotions dominate about COVID-19. It also examines how emotions and polarities change over time.

Figure 2 Eight emotions rates by date, (a) anger, (b) fear, (c) anticipation, (d) surprise, (e) disgust, (f) trust, (g) joy, (h) sadness, the red line shows the early period's rate and the blue line shows the later period's rate (see online version for colours)



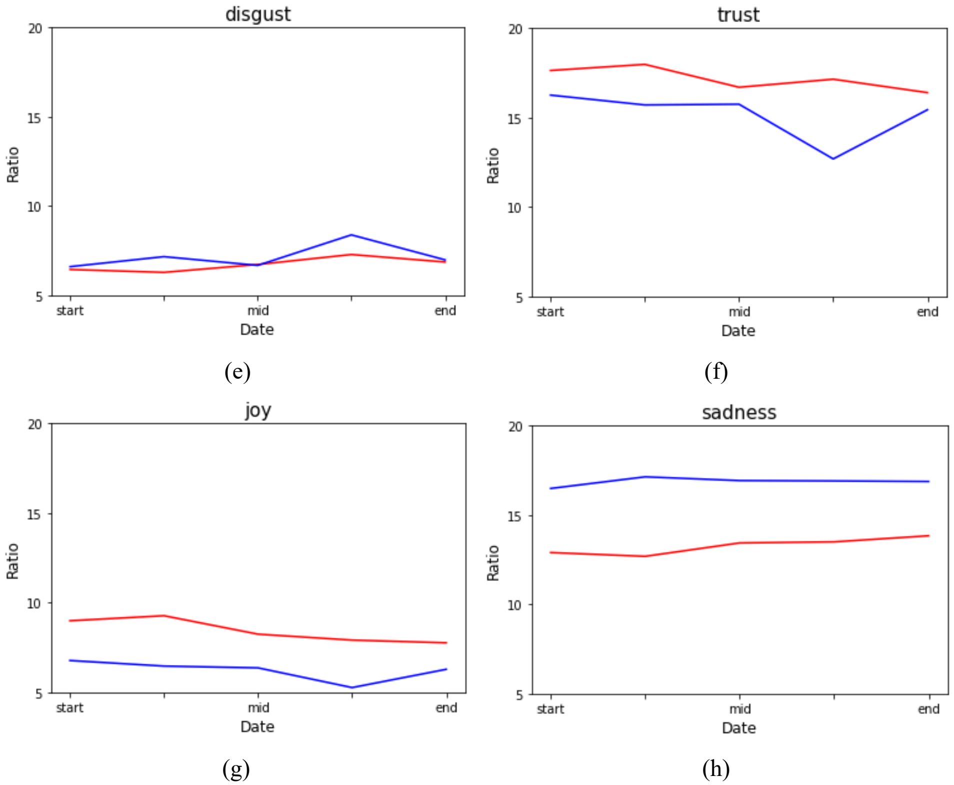
(a)

(b)

(c)

(d)

Figure 2 Eight emotions rates by date, (a) anger, (b) fear, (c) anticipation, (d) surprise, (e) disgust, (f) trust, (g) joy, (h) sadness, the red line shows the early period's rate and the blue line shows the later period's rate (continued) (see online version for colours)



(e)

(f)

(g)

(h)

4.1 National Research Council (NRC)

For sentiment analysis, the NRC Emotional Dictionary is used. It is labelled 0 and 1 for a total of 14,182 words, two polarities (positive, negative), and eight emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, trust). Since the ones labelled as 1 in the above dictionary are significant labels, the analysis is conducted in a way by extracting only them and counting the words contained in the tweet (Boon-Itt and Skunkan, 2020).

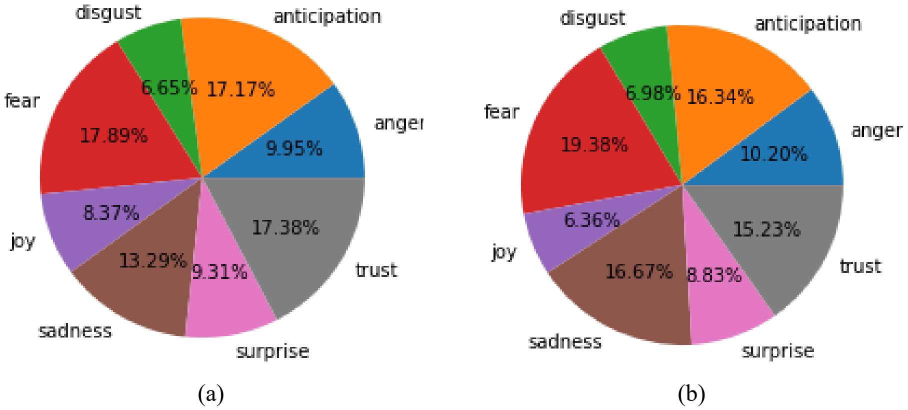
Summarising each period, the words are counted for eight emotions excluding two polarities, positives, and negatives. Then the ratio of the emotion is calculated and displayed in a graph among the total number of emotions per day, as shown in Figure 2.

According to Mohammad and Turney (2013), when performing the above NRC-based emotion analysis, there are pairs of emotions that are correlated with each other. First, looking at Figures 2(a) and 2(b), the anger is decreasing when the ratio of fear increases. In addition, the anger is increasing when the fear decreases, contrary to the previous one. Therefore, it can be confirmed that the two emotions of anger and fear are symmetrical to each other. In this way, Mohammad and Turney (2013) argue that anticipation-surprise, disgust-trust, and joy-sadness, also have a symmetrical relationship with each other, and the same result can be confirmed in Figures 2(c)–2(h).

In addition, looking at the ratio of emotions, Figure 3(a) shows that fear is the most common at 17.89% in the early period, trust is 17.38% and anticipation is 17.17%. In other words, fear about COVID-19 prevails, but the anticipation and trust that it will be resolved soon are supported.

On the other hand, looking at the later period of COVID-19, fear is 19.38%, the most dominant emotion as in the early period, and the ratio increased by about 1.5%. In addition, sadness is 16.67%, which is a significant number of dominant emotions following fear, and unlike the early period, the rate increased significantly to about 3.4% as COVID-19 lasted longer than expected. In the same context, the ratio of anger and disgust increased compared to the early period, but on the contrary, anticipation, trust, surprise, and joy decreased.

Figure 3 Eight emotions ratio, (a) the early period, (b) the later period (see online version for colours)

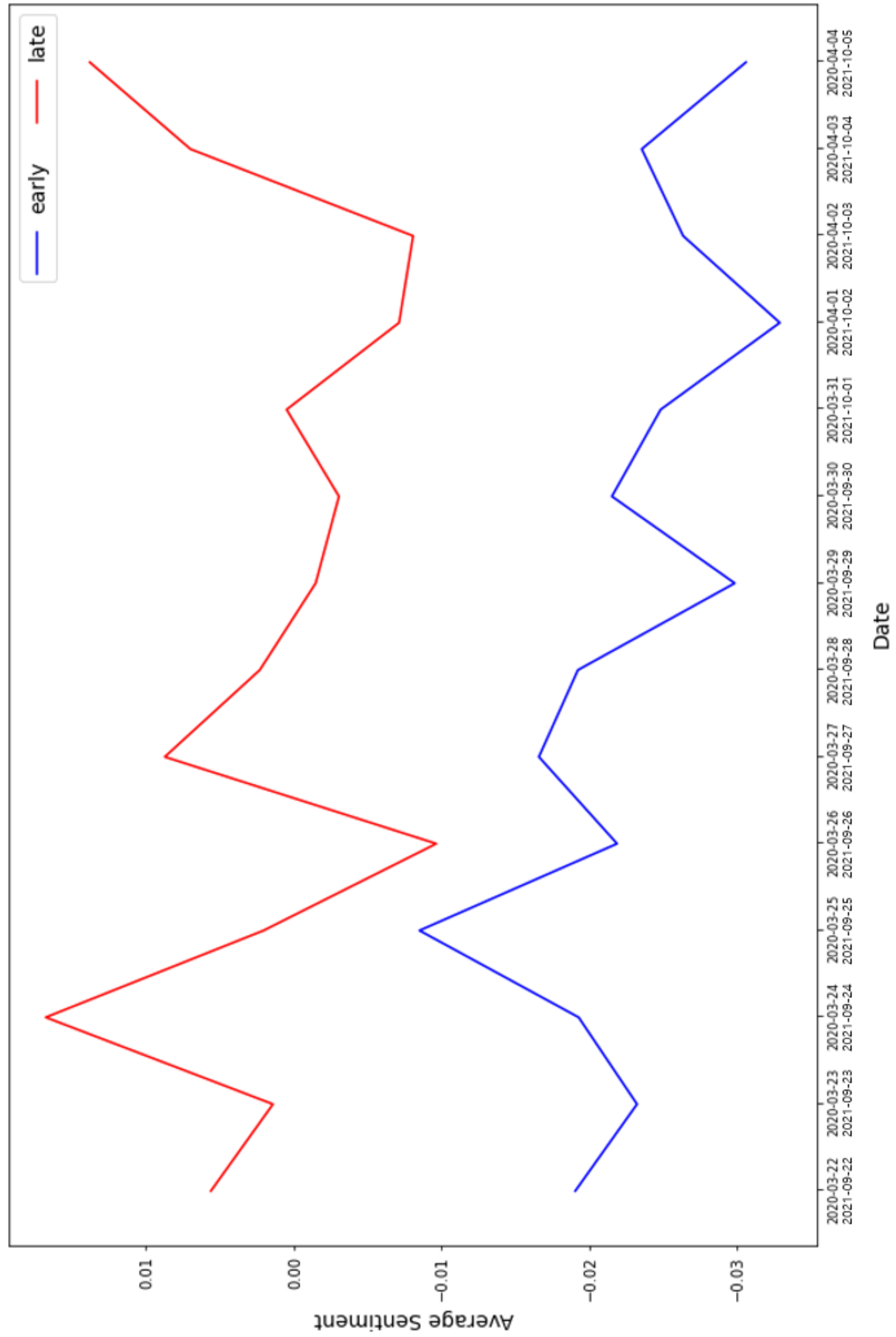


4.2 Sentimentr

The collected tweets are grouped into positive and negative tweets through emotion analysis conducted earlier (Chang and Wang, 2020), to find out what topics make up the emotion, and to compare and analyse the topics that make up the two emotions. Among the ten emotions that can be divided into emotion classifications for NRC, eighth emotions exclude two polarities, positive emotions include joy, trust, and negative emotions include anger, disgust, fear, and sadness (Mohammad and Turney, 2010). Anticipation and surprise can belong to both emotions. If classified as positive and negative according to this standard, negative emotions belong to far more emotions than positive emotions, so of course, negative tweets are included more. Therefore, different criteria are needed to classify positive and negative tweets.

To solve this problem, we calculate emotional scores using R's 'sentimentr' package. When the emotional polarity score is calculated for each tweet, the score is imposed with a value between –1 and 1 (Naldi, 2019) So tweets are classified by setting them to positive if it exceeds 0, negative if it is less than 0, and neutral if it is 0. As a result, 1,089,435 positive tweets, 1,229,137 negative tweets, and 544,270 neutral tweets are classified in the early period of COVID-19, and then 291,631 positive tweets, 265,022 negative tweets, and 133,262 neutral tweets are classified in the later period.

Figure 4 Emotional score average (see online version for colours)



Looking at the distribution by scoring emotions for each period of COVID-19, in the early case, the average of emotion scores for all dates is less than 0, so negative emotions are superior to positive emotions. In addition, as time passes, negative emotions gradually increase as the graph goes downward. On the contrary, in the latter case, the number of days when the emotional average scores greater than 0 is more than half at nine days, and it increases significantly compared to the early period. Through this, it can be confirmed that the public is changing positively about COVID-19 in the latter part, and the results are shown in Figure 4.

5 Topic analysis

We would like to find out what factors caused emotion using the results of the sentiment analysis conducted in the previous section. By identifying the main discussions of each emotion and period, it identifies people's general thoughts, including improvements. Therefore, Section 5 identifies the main causes of emotion induction by using LDA of topic modelling for topic analysis.

5.1 LDA

LDA, a representative method of topic modelling, is used to find out which topics constitute the emotion using datasets classified for positive and negative emotions. Tweets classified as neutral are removed, and experiments are conducted by classifying them into four datasets: positive early (PE), negative early (NE), positive late (PL), and negative late (NL). It is implemented using Python's 'Gensim' library, and it is also used for LDA model training by generating an id2word object in which words are stored one by one in a word dictionary to avoid overlapping words in the entire dataset (Chekijian et al., 2021).

5.2 Determining the number of topics

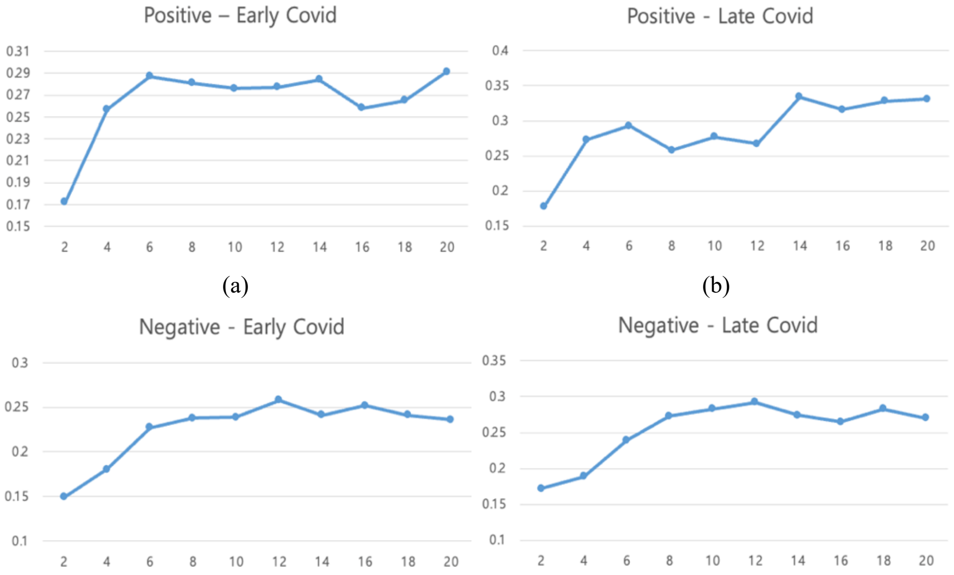
To effectively perform topic modelling, modelling is conducted by varying the number of topics for each emotion, and experiments are conducted to determine the optimal number of topics by comparing the results based on the coherence score. The values of the experimental results are shown in Figure 5.

In general, the coherence score increases as the number of topics increases, so the number of topics has the best score between 2 and 20, or the number of topics located at the inflection point is selected to determine the final 6. Accordingly, LDA's parameter, 'num_topics', proceeds to 6 as determined earlier as the number of topics as a result of the return of topic modelling. All other parameters such as 'passes' and 'iteration' proceed to the default value of the model.

5.3 Modelling results

Perplexity scores and coherence scores are used to evaluate whether the results of topic modelling have progressed well. Accordingly, topic modelling is performed on each of the four data sets classified according to emotion and period, and the results of the calculation are shown in Table 2.

Figure 5 Coherence score results according to the change in the number of topics, (a) PE dataset, (b) PL dataset, (c) NE dataset, (d) NL dataset (see online version for colours)



(a) (b)

(c) (d)

Table 2 Topic modelling calculation results

Score	*PE*	*PL*	*NE*	*NL*
Coherence (C_v)	0.261	0.355	0.226	0.265
Coherence (C_umass)	–4.429	–5.327	–4.517	–5.269
Perplexity	–8.618	–8.085	–8.744	–8.443

Table 3 PE dataset topic modelling result

#	*Topic words*	*Label*	*%*
1	Test, positive, response, news, first, two, patient, question, die	The current situation of COVID-19	21.5
2	Trump, new, come, post, world, American, vaccine, money, drug, news	Life changed due to COVID-19	18.3
3	New, case, death, confirm, York, update, total, home, stay, report	New York confirmed cases and deaths	18.2
4	People, good, time, hope, well, right, cure, great, take, stop	Hope to overcome COVID-19	15
5	Care, support, help, relief, health, donation, April, worker, protect, patient	How to cope with COVID-10	13.8
6	Mask, fight, keep, face, share, spread, safe, help, recommend, wear	COVID-19 response policy	13.2

For each dataset, check the topic, representative words, and proportions of the topic modelling results. First, the results of the PE dataset are shown in Table 3. As a result of directly generating the label through representative words, tweets about the current

COVID-19 situation are the most dominant. In addition, the life changed due to COVID-19, the number of confirmed cases and deaths in New York, and the topics of hope, method, and policy for overcoming COVID-19 are distributed.

In the PL dataset, COVID-19 statistics and worsening COVID-19 situations are dominant topics. Unlike the PE dataset result, Table 4 that new topics such as the impact, effects, and types of vaccines are emerging.

Table 4 PL dataset topic modelling result

#	*Topic words*	*Label*	*%*
1	New, case, death, report, update, record, latest, daily, confirm, active	COVID-19 statistics	21.3
2	Test, positive, rapid, unit, author, double, week, result, clinic, Pfizerbiontech	The worsening COVID-19 situation	18.6
3	Vaccine, die, new, mandatory, prevent, people, seek, school, live, York, vaccine, effect, study, show, Johnson	Influence of vaccines	16.7
4	Month, immune, support, protect, prevent	The effect of a vaccine	16.1
5	Vaccine, booster, Pfizer, dose, shot, receive, approve, FDA, fully, million	Types of vaccines	15.2
6	Job, impact, plan, emmi, relief, care, fund, teacher, celebrate, restrict	Everyday life was changed by COVID-19	12.1

The NE dataset is shown in Table 5. The spread of COVID-19, new kinds of news related to COVID-19, and government measures are the dominant topics. From the results below, the above topics are negatively recognised.

Table 5 NE dataset topic modelling result

#	*Topic words*	*Label*	*%*
1	Spread, help, stop, fight, mask, home, disease, risk, stay, let	The spread of COVID-19	17.9
2	Live, news, update, force, nurse, watch, crisis, war, new, task, crisis, lockdown, amid, fight, sign	New kinds of news update	17.9
3	Government, health, worker, family, state	The government's response	17.4
4	Death, die, case, news, toll, people, hospital, day, infect, patient	COVID-19 confirmed cases and deaths	16.7
5	People, think, cause, would, kill, fuck, come, really, time, vaccine	Opinions on COVID-19	15
6	Trump, test, response, china, drug, govern, warn, American, call, medic	COVID-19 situation in each country	15

Finally, the NL dataset is shown in Table 6. Unlike the early period, the type, effect, and influence of vaccines are emerging as new topics, and nevertheless, the number of COVID-19 confirmed cases, the global number of confirmed cases, and the risk of COVID-19 mutations are included, indicating why people's perception of vaccines is negatively included.

Table 6 NL dataset topic modelling result

#	*Topic words*	*Label*	*%*
1	Vaccine, shot, booster, emerge, Johnson, first, end, frustrate, grow, plan	Types and effects of vaccines	19
2	Vaccine, test, doctor, mandatory, school, Merck, refuse, Pfizer, news, govern	Influence of vaccines	19
3	Death, case, new, report, hospital, infect, pill, update, record, number	COVID-19 confirmed cases and deaths	17.4
4	Vaccine, flu, million, people, time, rate, world, sign, vax, immune	A global COVID-19 patient	15.5
5	Risk, fight, vaccine, health, effect, surgery, delta, world, drug, variant	The danger of COVID-19 mutation	15.4
6	Die, people, year, American, infect, day, man, study, treatment, start	COVID-19 related statistics	13.6

5.4 *Word cloud*

Based on the above topic modelling results, each calculated topic is expressed using a word cloud to find out what words are composed of Das and Dutta (2020). For each dataset, it is possible to check what keywords are composed of through Figure 6 in order. Since all four datasets have calculated six topics, keywords corresponding to topics 1, 2, and 3 above and topics 4, 5, and 6 below can be identified in order.

Figure 6 Topic modelling visualisation results, (a) PE dataset, (b) PL dataset, (c) NE dataset, (d) NL dataset (see online version for colours)

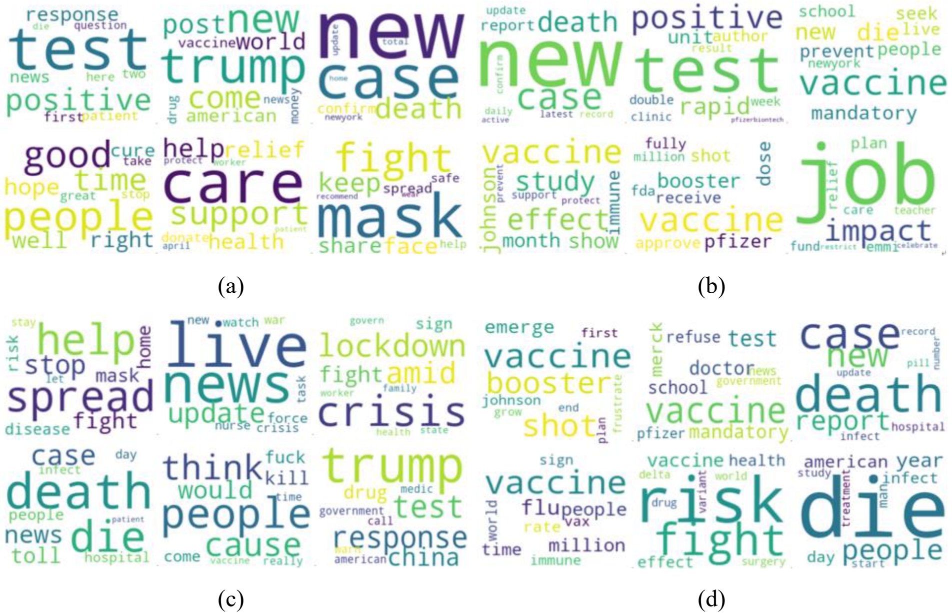
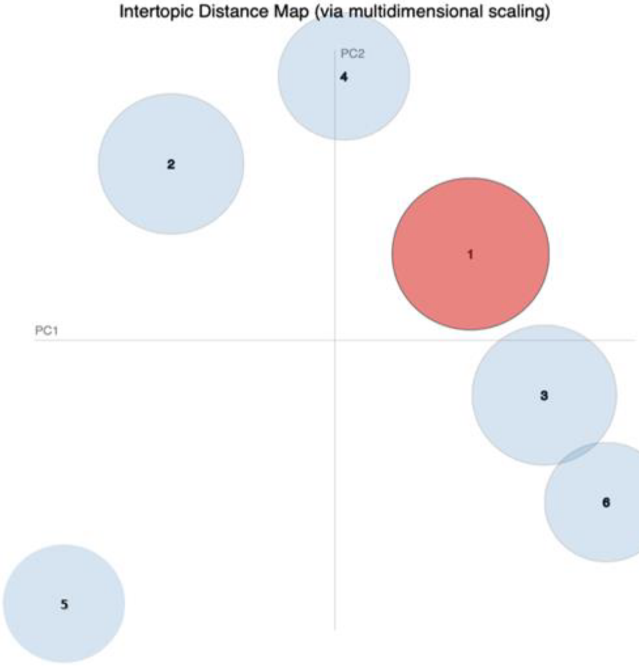
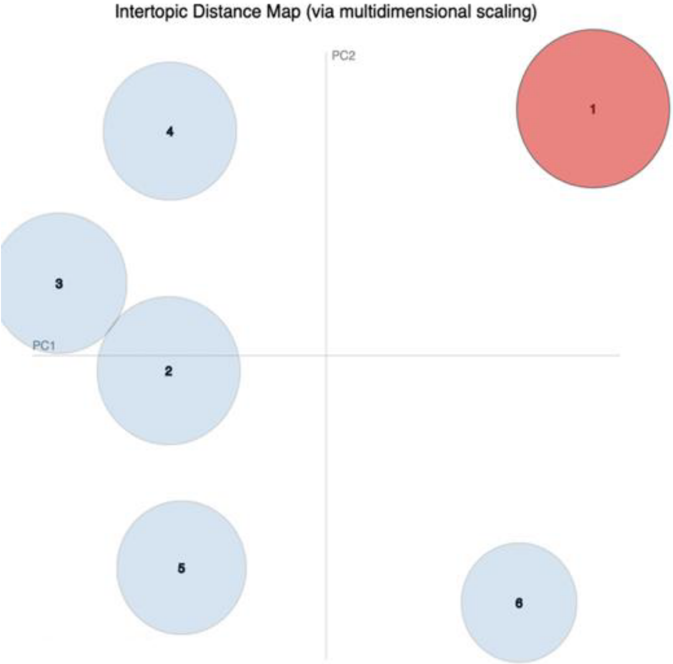

response test question die news here two positive first patient
post new vaccine world trump drug come news american money
new update total case home confirm death newyork
update report death new confirm case daily active latest record
positive unit author result test double clinic rapid week pfizerbiontech
school seek new die live prevent people newyork vaccine mandatory
good cure take hope time great stop people well right
help relief protect worker care support donate april patient health
fight keep safe recommend spread wear mask share face help
vaccine study prevent support protect johnson effect immune month show
fully million shot dose fda booster receive vaccine approve pfizer
plan job relief care teacher impact fund restrict emmi celebrate
(a)
(b)
stay help home risk stop let mask spread disease fight
new watch war live news task update force nurse crisis
govern sign lockdown fight amid worker family crisis health state
emerge first vaccine booster johnson grow end shot plan frustrate
merck refuse test doctor news school government vaccine pfizer mandatory
case record new update pill number death report hospital infect
case day infect death people patient news die toll hospital
think fuck kill would time people come cause vaccine really
trump drug medic government test call response american china
sign vaccine world flu people rate vax time million immune
vaccine health delta world drug risk variant fight effect surgery
american year study infect treatment man die day people start
(c)
(d)

Figure 7 Topic modelling visualisation results, (a) PE dataset, (b) PL dataset, (c) NE dataset, (d) NL dataset (see online version for colours)

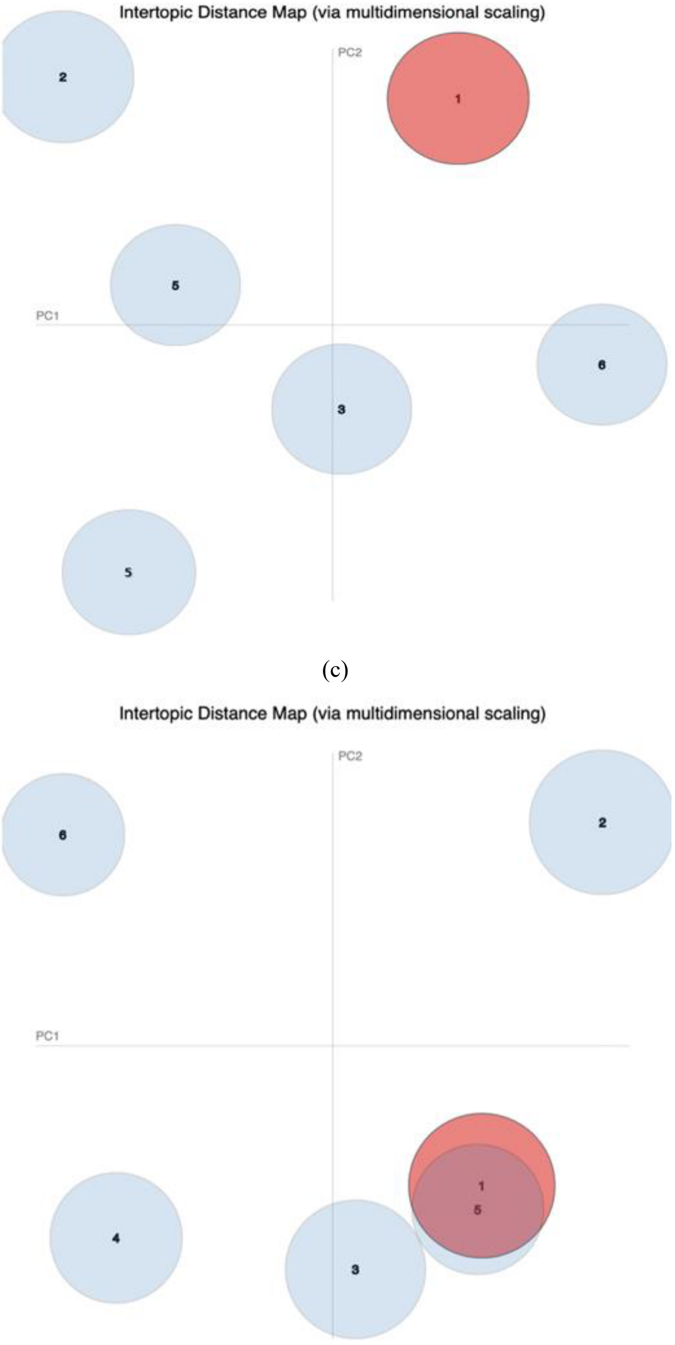


(a)



(b)

Figure 7 Topic modelling visualisation results, (a) PE dataset, (b) PL dataset, (c) NE dataset, (d) NL dataset (continued) (see online version for colours)



(c)

(d)

5.5 Visualisation of topic modelling

Based on the progressed topic modelling, the results of the LDA model are visually expressed using the 'pyLDAvis' library. The relationship between topics and topic keywords is easily understood by using the dimension reduction method, principal component analysis (PCA), and the keyword extraction method. The topic has the dimension of the number of words, so the dimension reduction method is used to compress it into two dimensions for visualisation. After extracting the information necessary for visualisation, it is converted into JSON and rendered on an HTML page, and the visualisation result is shown in Figure 7.

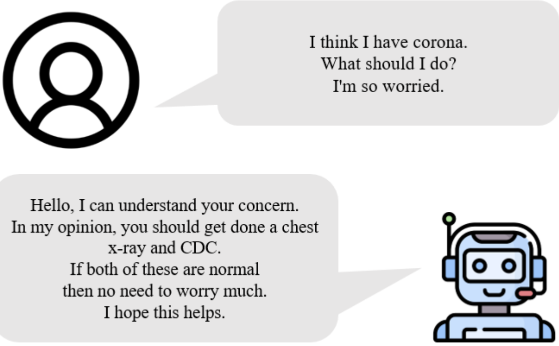
This is an 'inter-topic distance map', where the size of the circle indicates how many words in the topic belong and how they are distributed (Hidayatullah et al., 2018). In addition, the distance between circles is the similarity between topics, for example, if two circles overlap, the two topics are similar. In Figure 6(a), topic 1 contains more words than other topics, and in Figure 6(d), topics 1 and 5 have higher similarities than other topics.

6 COVID-chatbot

As COVID-19 continues to spread, a system is needed to answer any ongoing questions. In addition, through sentiment analysis, despite the emergence of vaccines, negative emotions still exist in the constant COVID-19 situation, and emotions such as fear, sadness, and disgust are increasing. Therefore, services for treatment for people at risk of depression such as 'Corona Blue' are needed.

Based on the analysis results of this study, as shown in Figure 8, we implement a COVID-chatbot to provide a question-and-answer service and a service that provides counselling and treatment by recognising emotions.

Figure 8 Motivating example (see online version for colours)



I think I have corona.
What should I do?
I'm so worried.

Hello, I can understand your concern.
In my opinion, you should get done a chest x-ray and CDC.
If both of these are normal
then no need to worry much.
I hope this helps.

6.1 Chat-bot dataset

We train COVID-chatbot using dialogue datasets of people's questions and answers related to COVID-19 and medical information. For COVID-chatbot training, frequently asked questions and responses related to COVID-19 were collected by CDC, CNN, GitHub, FDA, Illinois Department of Public Health (IDPH), John Hopkins Medicine, UN, WJLA, etc. (Jerry et al., 2020), and the valid test set that specialised in consultation with doctors regarding COVID-19 were used (Wenmian et al., 2020). In addition, to train the chatbot to answer basic medical knowledge, the COVID-chatbot is trained using a test set containing pairs of medical questions and answers such as treatment, diagnosis, and side effects of various diseases (reference 추가!!). To generate and train the appropriate topic and sentimental labels through experiments, other additional information contained in the dataset is removed, and only question and answer data are used. As the result, training is carried out with a total of 3,086 datasets, both of which are combined.

6.2 Topic label with LDA_BERT

This paper first proceeds with topic classification to generate a topic label suitable for the corresponding dataset. At this time, to determine the optimal number of topics, the change in the coherence score of the LDA model according to the change in the number of topics is shown in Figure 9. In general, the coherence score of the modelling tends to increase as the number of topics increases, so the number of topics is designated as 7 through the following results.

Figure 9 Coherence score results according to the number of topics (see online version for colours)

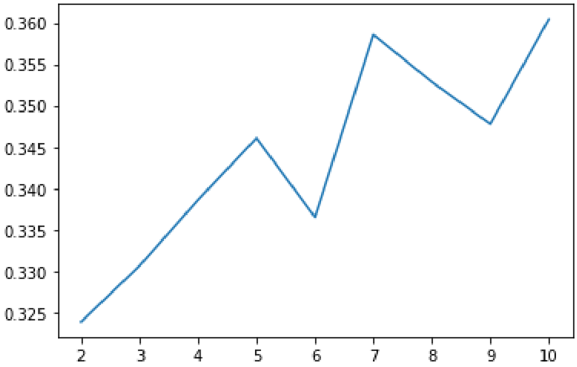


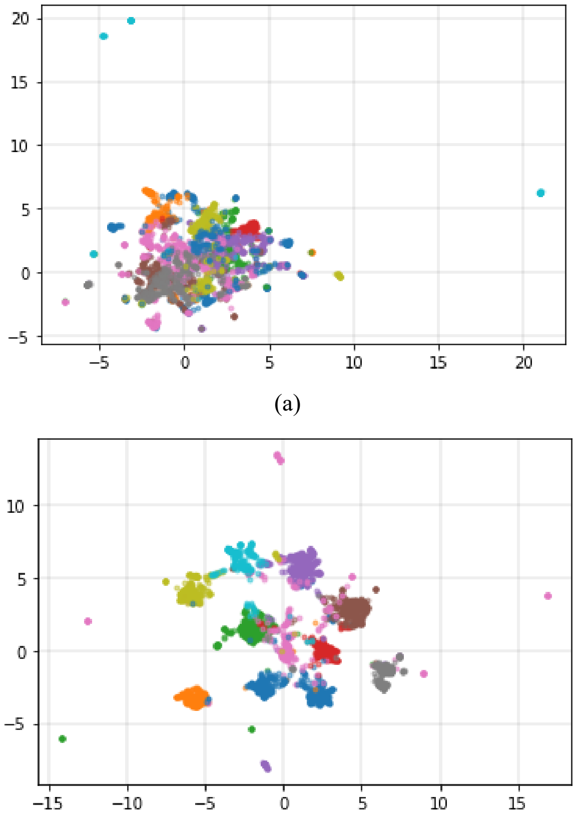
Table 7 Topic classification experiment results

Score	*LDA*	*BERT*	*LDA_BERT*
Coherence	0.3417	0.4603	0.4774
Silhouette	None	0.0523	0.2582

To classify the topics of the chatbot conversation dataset, we intend to create a topical label by using the LDA of the topic modelling discussed earlier and BERT, a language model that exhibits excellent performance in classification. Balance the relative importance of information by linking probabilistic subject assignment vectors in LDA and sentence embedding vectors in BERT with weighted hyperparameters. We also use a nonlinear transformation, auto encoder, to learn sub-dimensional latent spatial representations of connected vectors. Therefore, classify topics through methods of designating topics through clustering of latent space representations (Maarten, 2022; Steve, 2020; Nicole et al., 2020). Table 7 shows the modelling results of classifying topics using LDA and BERT together.

The 'LDA_BERT' method shows that in the case of coherence score, the performance is about 0.14 and 0.02 higher than that of executing for each of LDA and BERT. In the case of silhouette core, it is possible to confirm excellent results by about four times more when used with LDA compared to BERT. When classifying topics using LDA and BERT together, both coherence scores and silhouette scores perform better than classifying topics using only LDA and BERT alone.

Figure 10 Modelling visualisation results, (a) BERT, (b) LDA_BERT (see online version for colours)



(a)

()

Furthermore, looking at the results by visualising the results of the modelling in 2D embeddings, from Figure 10(a) and 10(b) that the subject is classified more cohesively when used with LDA than when BERT was used alone.

6.3 Sentiment label

It is trained through the sentimental label and BERT of the dataset to generate answers that match the emotions of the text entered in the COVID-chatbot (Manish et al., 2019). At this time, the dataset used for chatbot training does not have a label with emotions classified, so a process of generating it is necessary. The emotion score is calculated using the rule-based emotion analysis tool 'VADER' in the nltk library (Shihab and Yang, 2019). If the value is 0, it is classified as neutral, if it is greater than 0, it is classified as positive, and if it is less than 0, it is classified as negative to generate an emotional label. Looking at the ratio is shown in Figure 11.

Figure 11 Distribution of emotion classification (see online version for colours)

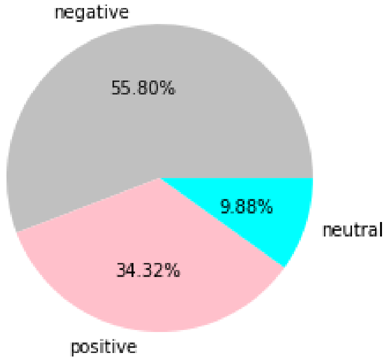


Table 8 BERT sentiment classification results

Sentiment	*Precision*	*Recall*	*F1-score*
Negative	0.87	0.88	0.88
Neutral	0.74	0.74	0.74
Positive	0.79	0.77	0.78

And also emotion classification is conducted with BERT, and experiments are conducted to find out the performance of the model. The experimental results are shown in Table 8, showing excellent classification performance.

6.4 COVID-chatbot

This paper seeks to utilise a pre-trained language model for the implementation of a chatbot. Among GPT-2, which shows high performance in sentence generation provided by Microsoft, we proceed with implementation using DialoGPT-small, which is specialised for conversation. The COVID-chatbot utilises the topic labels and sentimental labels generated to identify the emotions included in the content along with the topic of conversation about the user's input and to generate the appropriate answer accordingly (Peishu et al., 2020). Fine-tuning of the language model is performed by adding each label as a special token along with the training data.

As a measure of evaluating the predictive answers produced by the implemented chatbot, it is evaluated using standard machine translation metrics such as Perplexity, Meteor, NIST, etc. (Zhang et al., 2019). Perplexity is a value calculated based on the cumulative probability of selected tokens, assuming that one token with the greatest probability is selected among thousands of tokens when generating a sentence, and it is judged that the lower the perplexity value, the better the learning. Meteor is an evaluation method derived from BLEU, which complements the incompleteness of BLEU handling only precision and considers recall together, and it means that the higher the result value along with the NIST value, the better the performance.

Table 9 Chat-bot experimental results

	Perplexity	*Meteor*	*NIST_2*	*NIST_4*
Baseline	76.0903	0.9696	0.2307	0.2342
SENT label	64.3090	0.1249	0.3079	0.3144
TOPIC label	72.4983	0.1263	0.3908	0.3971
SENT label + TOPIC label	56.8567	0.1340	0.4936	0.5043

The results of the experiment are shown in Table 9. This confirms that when trained together using the topic label and sentimental label generated as a token rather than simply training the chatbot with a dataset of questions and answers, all evaluation scales produce better answers. As a result, it can be seen that compared to a baseline without labels, it shows a higher performance when training by adding sentimental and topical labels, and when both labels are added, it produces a much more natural and appropriate answer.

7 Conclusions

Social media data analysis can explain the public feelings and perceptions of COVID-19, and this study deals with discussions on COVID-19 written on Twitter and investigates related emotions and topics. For effective analysis, the analysis was conducted by dividing it into four categories based on emotion and period. As a result of sentiment analysis, negative emotions were overwhelming in the early period when the COVID-19 epidemic began, under the themes of the spread of COVID-19, the rapidly increasing number of confirmed cases and deaths, and life controlled by COVID-19. On the other hand, it was confirmed that in the later period, a year, and a half later, the subjects of the COVID-19 situation, which improved with the appearance of the vaccine, were gradually changed into positive emotions. However, even though vaccines still exist, there are also negative feelings about COVID-19, which does not end with the emergence of a new mutant virus.

Corona is a very new disease, and many things are still unknown. This uncertainty has led people to find and share information, and to this end, they mainly use social media. Through the topic analysis conducted, efficient information related to COVID-19 is still insufficient as topics such as the current situation of COVID-19, daily life changed to COVID-19, policies, and types and effects of vaccines occupy a large proportion.

To solve the curiosity and depression triggered by COVID-19, the COVID-chatbot is trained using the dialogue dataset with questions and answers related to COVID-19 and the consultation dataset with a doctor. However, since COVID is a relatively recent disease that occurred at the end of 2019, there are not many related QA data sets, and the number of data constituting is also very small. Therefore, although there is a limitation that the result value of the experiment conducted in this paper is relatively low, it can be supplemented by learning additional datasets. Nevertheless, by creating a topic label and a sentimental label suitable for the corresponding dataset and then training DialoGPT with the special token, the COVID-chatbot was implemented to generate answers suitable for the topic and emotion of the user-written input text. Therefore, necessary information related to COVID-19 can be obtained through the implemented chatbot, and people's emotions can be comforted and treated.

In future studies, based on the analysis results of this study, the emotional chatbot will be advanced in consideration of big graphs. It also wants to expand chatbots to other areas, following the trend that COVID-19 is ending.

Acknowledgements

This research was supported by the Basic Science Research Program through the NRF (National Research Foundation of Korea), funded by the MSIT (Ministry of Science and ICT), and the Gachon University research fund of 2021 (Nos. NRF2019R1A2C1008412, GCU-202103390001).

References

Banda, J.M., Tekumalla, R., Wang, G., Yu, J., Liu, T., Ding, Y., Artemova, E., Tutubalina, E. and Chowell, G. (2021) 'A large-scale COVID-19 Twitter chatter dataset for open scientific research – an international collaboration', *Epidemiologia*, Vol. 2, No. 3, pp.315–324, https://doi.org/10.3390/epidemiologia2030024.

Boon-Itt, S. and Skunkan, Y. (2020) 'Public perception of the COVID-19 pandemic on Twitter: sentiment analysis and topic modeling study', *JMIR Public Health and Surveillance*, Vol. 6, No. 4, https://doi.org/10.2196/21978.

Chang, C. and Wang, X. (2020) 'Research on dynamic political sentiment polarity analysis of specific group Twitter based on deep learning method', *Journal of Physics: Conference Series*, https://doi.org/10.1088/1742-6596/1651/1/012108.

Chekijian, S., Li, H. and Fodeh, S. (2021) 'Emergency care and the patient experience: using sentiment analysis and topic modeling to understand the impact of the COVID-19 pandemic', *Health and Technology*, Vol. 11, No. 5, pp.1073–1082, https://doi.org/10.1007/s12553-021-00585-z.

Das, S. and Dutta, A. (2020) 'Characterizing public emotions and sentiments in COVID-19 environment: a case study of India', *Journal of Human Behavior in the Social Environment*, pp.154–167, https://doi.org/10.1080/10911359.2020.1781015.

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2018) *Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding*, arXiv preprint, arXiv: 1810.04805.

Heo, S. and Yang, J. (2020) 'Analysis of research topics and trends on COVID-19 in Korea using latent Dirichlet allocation (LDA)', *Journal of the Korea Society of Computer and Information.*, Vol. 25, No. 12, pp. 83–91, https://doi.org/10.9708/JKSCI.2020.25.12.083.

Hidayatullah, A.F., Pembrani, E.C., Kurniawan, W., Akbar, G. and Pranata, R. (2018) 'Twitter topic modeling on football news', *2018 3rd International Conference on Computer and Communication Systems (ICCCS)*, pp.467–471, https://doi.org/10.1109/CCOMS.2018.8463231.

Hussain, A., Tahir, A., Hussain, Z., Sheikh, Z., Gogate, M., Dashtipour, K. and Sheikh, A. (2021) 'Artificial intelligence-enabled analysis of public attitudes on Facebook and twitter toward COVID-19 vaccines in the United Kingdom and the United States: observational study', *Journal of Medical Internet Research*, Vol. 23, No. 4, https://doi.org/10.2196/26627.

Jerry, W., Chengyu, H., Soroush, V. and Jason, W. (2020) 'What are people asking about COVID-19? A question classification dataset', in *NLP for COVID-19 Workshop at ACL*, https://doi.org/10.48550/arXiv.2005.12522.

Kim, M., Ryu, J., Cha, D. and Sim, M. (2020) 'Stock price prediction using sentiment analysis: from 'stock discussion room in Naver', *The Journal of Society for e-Business Studies*, Vol. 25, No. 4, pp.61–75, https://doi.org/10.7838/jsebs.2020.25.4.061.

Kim, N., Lee, D., Choi, H. and Wong, W.X.S. (2017) 'Investigations on techniques and applications of text analytics', *The Journal of Korean Institute of Communications and Information Sciences*, Vol. 42, No. 2, Korea Information and Communications Society, https://doi.org/10.7840/kics.2017.42.2.471.

Lyu, J.C., Han, E.L. and Luli, G.K. (2021) 'COVID-19 vaccine-related discussion on Twitter: topic modeling and sentiment analysis', *Journal of Medical Internet Research*, Vol. 23, No. 6, https://doi.org/10.2196/24435.

Maarten, G. (2022) *BERTopic: Neural Topic Modelling with a Class-Based TF-IDF Procedure*, https://doi.org/10.48550/arXiv.2203.05794.

Manguri, K.H., Ramadhan, R.N. and Mohammed Amin, P.R. (2020) 'Twitter sentiment analysis on worldwide COVID-19 outbreaks', *Kurdistan Journal of Applied Research*, Vol. 5, No. 3, pp.54–65, https://doi.org/10.24017/covid.8.

Manish, M., Sushil, S. and Aakash, S. (2019) 'Fine-grained sentiment classification using BERT', *Artificial Intelligence for Transforming Business and Society (AITB)*, pp.1–5 [online] http://arxiv.org/abs/1910.03474.

Mehri, S. and Eskenazi, M. (2020) *Unsupervised Evaluation of Interactive Dialog with dialogpt*, arXiv preprint arXiv: 2006.12719.

Mohammad, S.M. and Turney, P.D. (2010) *Emotions Evoked by Common Words and Phrases: using Mechanical Turk to Create an Emotion Lexicon*, Association for Computational Linguistics, pp.28–34 [online] https://dl.acm.org/doi/10.5555/1860631.1860635 (accessed 5 June 2010).

Mohammad, S.M. and Turney, P.D. (2013) 'Crowdsourcing a word-emotion association lexicon', *Computational Intelligence*, Vol. 29 [online] http://arxiv.org/abs/1308.6297 (accessed 4 October 2019).

Naldi, M. (2019) *A Review of Sentiment Computation Methods with R Packages* [online] http://arxiv.org/abs/1901.08319 (accessed 24 January 2019).

Nhung, D., Wenny, R. and Torab, T. (2016) 'A query expansion approach for social media data extraction', *International Journal of Web and Grid Services (IJWGS)*, Vol. 12, No. 4, pp.418–441, https://doi.org/10.1504/IJWGS.2016.080142.

Nicole, P., Dong, N. and Maria, L. (2020) 'tBERT: topic models and BERT joining forces for semantic similarity detection', in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp.7047–7055 [online] https://aclanthology.org/2020.acl-main.630 (accessed 16 January 2022).

Park, S., Do, K., Kim, H., Park, G., Yun, J. and Kim, K. (2018) 'An exploratory study of happiness and unhappiness among Koreans based on text mining techniques', *The Journal of the Korea Contents Association*, Vol. 18, No. 7, pp.10–27, https://doi.org/10.5392/JKCA.2018.18.07.010.

Peishu, H., Yang, Y., Zhou, J., Chen, C. and He, L. (2020) 'TERG: topic-aware emotional response generation for chatbot', *2020 International Joint Conference on Neural Networks (IJCNN)*, pp.1–8, https://doi.org/10.1109/jksci.2020.25.12.083.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I. (2019) 'Language models are unsupervised multitask learners', *OpenAI blog*, Vol. 1, No. 8 [online] https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_ unsupervised_multitask_learners.pdf (accessed 3 December 2020).

Shihab, E. and Yang, J. (2019) 'Twitter sentiment analysis using natural language toolkit and VADER sentiment', *Proceedings of the International Multiconference of Engineers and Computer Scientists (IMECS)*, Vol. 122, p.16 [online] http://www.iaeng.org/publication/ IMECS2019/IMECS2019_pp12-16.pdf (accessed 13–15 March 2019).

Steve, S. (2020) *Contextual Topic Identification: Identifying Meaningful Topics for Sparse Steam Reviews* [online] https://medium.com/insight-data/contextual-topic-identification-4291d256a032 (accessed 5 March 2020).

Sun, X., Liu, X., Hu, J. and Zhu, J. (2014) *Empirical Studies on the NLP Techniques for Source Code Data Preprocessing*, Association for Computing Machinery, pp.32–39, https://doi.org/ 10.1145/2627508.2627514.

Tyagi, P. and Tyripathi, R.C. (2019) 'A review towards the sentiment analysis techniques for the analysis of Twitter data', *Proceedings of 2nd International Conference on Advanced Computing and Software Engineering (ICACSE)*, http://dx.doi.org/10.2139/ssrn.3349569.

Wenmian, Y., Guangtao, Z., Bowen, T., Zeqian, J., Subrato, C., Xuehai, H., Shu, C., Xingyi, Y., Qingyang, W., Zhou, Y., Eric, P.X. and Pengtao, X. (2020) *On the Generation of Medical Dialogues for COVID-19*,https://doi.org/10.48550/arxiv.2005.05442.

Williams, E.M., Levin, D. and McCulloh, I. (2020) 'Improving LDA topic modeling with Gamma and Simmelian filtration', *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp.692–696, https://doi.org/10.1109/ ASONAM49781.2020.9381330.

Xue, J., Chen, J., Hu, R., Chen, C., Zheng, C., Su, Y. and Zhu, T. (2020) 'Twitter discussions and emotions about the COVID-19 pandemic: machine learning approach', *Journal of Medical Internet Research*, Vol. 22, No. 11, p.1, https://doi.org/10.2196/20550.

Yoo, S., Kim, D., Yang, S. and Jeong, O. (2020) 'Real-time disease detection and analysis system using social media contents', *International Journal of Web and Grid Services (IJWGS)*, Vol. 16, No. 1, pp.22–38, https://doi.org/10.1504/IJWGS.2020.106103.

Yousefinaghani, S., Dara, R., Mubareka, S., Papadopoulos, A. and Sharif, S. (2021) 'An analysis of COVID-19 vaccine sentiments and opinions on Twitter', *International Journal of Infectious Diseases*, Vol. 108, pp.256–262, https://doi.org/10.1016/j.ijid.2021.05.059.

Yun, Y., Jo, J., Hur, Y. and Lim, H. (2017) 'A comparative analysis of cognitive change about big data using social media data analysis', *KIPS Transactions on Software and Data Engineering*, Vol. 6, No. 7, pp.371–378, https://doi.org/10.3745/KTSDE.2017.6.7.371.

Zhang, Y., Sun, S., Galley, M., Chen, Y.C., Brockett, C., Gao, X., Gao, J., Liu, J. and Dolan, B. (2019) *Dialogpt: Large-Scale Generative Pre-Training for Conversational Response Generation*, arXiv preprint arXiv: 1911.00536.

Zoya, Latif, S., Shafait, F. and Latif, R. (2021) 'Analyzing LDA and NMF topic models for Urdu tweets via automatic labeling', in *IEEE Access*, Vol. 9, pp.127531–127547, https://doi.org/ 10.1109/ACCESS.2021.3112620.