Preprint submitted to the Elsevier Computers & Electrical Engineering Journal. You can find the published version at:
https://www.sciencedirect.com/science/article/pii/S0045790622007765?dgcid=author

Detecting and Analyzing Topics of Massive COVID-19-Related Tweets for Various Countries

Faezeh Azizi
Perlab, Faculty of Electrical and Computer Engineering
University of Birjand
Birjand, Iran
faezeh.azizi1995@birjand.ac.ir

Hamed Vahdat-Nejad*
Perlab, Faculty of Electrical and Computer Engineering
University of Birjand
Birjand, Iran- Corresponding author
vahdatnejad@birjand.ac.ir

Hamideh Hajiabadi
Department of Computer Engineering
Birjand University of Technology
Birjand, Iran
hajiabadi@birjandut.ac.ir

Mohammad Hossein Khosravi
Faculty of Electrical and Computer Engineering
University of Birjand
Birjand, Iran
mohokhosravi@birjand.ac.ir

***Abstract*— with the flare-up of the COVID-19 infection since 2020, COVID-19 has been one of the hottest topics on Twitter. Topic modeling is one of the most popular analyzes, which extracts the topics from the text. This paper proposes a method to extract the most-discussed topics for 32 countries of the world. In this regard, more than five million related tweets have been studied and a method based on content analysis is proposed to identify the exact location of each tweet. Then, by using the statistical algorithm of Latent Dirichlet Allocation, the main topics of the tweets are identified. By leveraging sentiment analysis, the topics are afterward divided into positive and negative groups, and their trends in a quarterly period are investigated for the countries under study. The outcome of the analysis of time trends shows that for most countries, the trend of negative topics is highly correlated with the number of confirmed cases of COVID-19.**

Keywords—Topic modeling, Twitter social network, LDA, Sentiment analysis, COVID-19

I. INTRODUCTION

The coronavirus quickly became a global epidemic in 2020. According to Worldometer[1], about 180 million people worldwide were infected with the coronavirus and almost four million people died because of the virus in 2020. Besides, the coronavirus has been one of the hottest topics of research, so that in 2022, the search for the keyword COVID-19 in the Google Scholar[2] database leads to the retrieval of more than 120,000 articles in various fields.

Users of social media such as Twitter provide a rich resource of information by sharing their opinions. This information can be analyzed by researchers to extract knowledge. Twitter has influenced not only the media but also the discovery of breaking news and the news production cycle [1]. Twitter users have paid close attention to the COVID-19 virus during the outbreak and global quarantine, with COVID-19 being the most-discussed topic on Twitter in 2020[3] where more than 400 million tweets have been shared about it. Numerous researchers have analyzed and reviewed these tweets using supervised and unsupervised machine learning methods for different purposes [2-5]. Supervised methods cannot be used much in processing tweets, because tweets usually do not have a class label, and manually labeling is time-consuming and practically very difficult; therefore, leveraging unsupervised methods to analyze tweets is more common.

One of the unsupervised methods to analyze user tweets is topic modeling [6], which is one of the most popular methods to get a summary of the top topics. Previously, detecting topics by topic modeling in Twitter has been considered in the context of diseases and epidemics. For example, exploration of the topics regarding the effectiveness of the drugs used for combat seasonal flu [7], as well as the topics that users discussed during the Zika epidemic in the Americas [1]. Furthermore, a few research studies have been conducted to identify the top topics of COVID-19 tweets for a limited number of countries [8-10]. In this study, we analyze a large number of published tweets amid the early months of the corona outbreak. To accurately detect the location of each tweet, a content analysis method [9] is proposed to identify the location of each tweet. We then identify the hot topics in the 32 countries which are most affected by the Coronavirus using the Latent Dirichlet Allocation (LDA) [11] topic modeling algorithm. The proposed method is an extension to the previous article [9], which deals only with the topics of four countries. This article extends the previous research by examining the topics of tweets as well as identifying positive and negative topics and cumulative trends for 32 countries.

The proposed method is implemented in the GATE [12] tool on a database containing more than 5 million tweets. It results in detecting ten topics including "Reopening", "Death cases", "Telecommuting", "Protests", "Anger expression", "Masking", "Medication", "Social Distancing", "Second Wave" and "Peak of the disease" [9]. The main questions that are examined in this research are: (i) In which countries people have mainly positive attitude towards COVID-19 and in which countries there are mainly negative thoughts (ii) What are the most-discussed topics and how frequently is each topic

[1] https://www.worldometers.info/coronavirus/
[2] https://scholar.google.com
[3] https://blog.twitter.com/en_us/topics/insights/2020/spending-2020-together-on-twitter

discussed compared to other topics? (iii) How does the trend of positive and negative topics change in each country over the studied time, and are these changes correlated with the confirmed cases in that country?

To the best of our knowledge, this paper is the primary work in identifying topics and sentiment analysis about tweets for 32 countries infected with the COVID-19 virus. In addition, we analyze the time trend of negative and positive topics for each country.

The rest of the article is set as follows: Section 2 presents related works that have used topic modeling to analyze COVID-19-related tweets. Section 3 describes the proposed method and section 4 deals with the implementation and analysis of the proposed method. Section 5 discusses the conclusions and limitations of the research.

II. Related Work

As an epidemic, COVID-19 is the most important issue that has severely affected human life since 2020. The most used Twitter hashtag in 2020 was dedicated to # covid19 and users have shared more than 400 million tweets with this hashtag[4].

People generally share their views, preferences, and thoughts about various topics through social media such as Facebook and Twitter publicly. There is some information previously extracted by analyzing the users' reviews about COVID-19 on social media; for example, “fear of the virus has been the dominant sentiment” among Twitter users [13].

One of the most interesting analyzes and processes on people's opinions is topic modeling. With the help of this processing, the key and hot topics of users' discussions in the context of cyberspace can be detected. Recently, topic extraction with the help of topic modeling on Twitter has gotten attention in the field of diseases and epidemics, for example, collecting experiences of Twitter users regarding the effectiveness of drugs to combat seasonal flu [14] and collecting Twitter users' discussions on the Zika epidemic in the Americas [15].

Amid the spread of the COVID-19 epidemic, several studies have leveraged topic modeling for tweet processing. Finding the topics of tweets during this period, they have been looking for items such as analyzing users' loneliness in tweets [18], users' attitudes about domestic violence [19], news orientation of Spanish newspapers accounts [20], users' attitudes about gambling [21], the impact of COVID-19 on the stock market [22], people's attitudes towards the effect of climate on the spread of the virus [4], people's attitudes about using web-conferencing systems [23], and the discussions on vaccination [24]. Moreover, a few research studies have identified topics and analyzed the sentiments of tweets for the whole world [13][25-27].

Several research studies have investigated the topics of tweets in specific countries. In this regard, Klaifer and Berton [28] identified the top topics of COVID-19 related tweets for both Brazil and the United States. They collected nearly 7,000 tweets in English and Portuguese, published from April 17 to August 8, 2020. Users' addresses have been used to obtain the geographical location of these countries. They extracted ten topics for English-language tweets and ten topics for Portuguese-language tweets using the Gibbs Sampling algorithm for the Dirichlet Multinomial Mixture (GSDMM). Topics of English Tweets have the titles of "Economic impacts", "Case reports / statistics", "Proliferation care", "Politics", "Entertainment", "Treatments", "Online events", "Charity", "Sports", and "Anti-racism protests" [28].

Doogan et al. [8] have extracted the top topics of COVID-19 tweets for six countries including Australia, Canada, New Zealand, Ireland, the United Kingdom, and the USA. They used hashtags to obtain the geographical location of countries, and collected 700,000 tweets from January 1 to April 30, 2020, and used the MetaLDA method to identify topics. For this purpose, 130 topics have been identified, and among them, topics related to non-pharmacological interventions have been isolated. Then, the topics are divided into two categories including non-drug restrictive interventions (“Lockdowns”, “Gathering restrictions”, “Travel” “Restrictions”, and “Workplace closures”) and restrictive drug interventions ("Personal Protection", "Social Distancing", and "Testing and Tracing") [8]. Finally, the ratio of topics discussed for each country has been determined, with New Zealand receiving the most attention and the USA the least attention to drug interventions.

Jang and colleagues [10] have extracted the top topics of COVID-19 related tweets for both Canada and the USA. They used the geotag location of tweets to collect the tweets related to these countries. For this purpose, 300,000 shared tweets were collected from February 15 to March 31, and the LDA method was used to identify topics. Then 20 topics were identified for these two countries and categorized into "public health promotions" and "interventions"[10].

Finally, in our previous work [9], we identified the hot topics of COVID-19-related tweets from March 23 to June 23, 2020. For this purpose, we have used the LDA algorithm to identify the topics and then the topics with the titles "reopening", "death cases", "telecommuting", "protests", "anger expression", "masking", "medication", "social distance", "second wave", and "peak of the disease" have been extracted [9]. We also have analyzed time trends for four countries including China, Canada, the USA, and the UK. In this article, we extend the previous work by investigating 32 countries with the highest COVID-19 cases as well as published tweets. We use sentiment analysis to identify the negative and positive-thinking countries. Taking into account the quarterly period, we examine the trend of negative and positive topics for each country to analyze the fluctuations of the topics and their relationship with the confirmed cases of COVID-19.

III. Proposed Method

The proposed method includes the main operations of collecting tweets, extracting the location, preprocessing, analyzing sentiment, and topic modeling (Fig.1), which are described in this section. According to blog.twitter.com[5], the top Twitter hashtag in 2020 belongs to "COVID-19" and the third top Twitter hashtag in 2020 belongs to the hashtag "Stay

4 https://blog.twitter.com/en_us/topics/insights/2020/spending-2020-together-on-twitter.html

5 https://blog.twitter.com/en_us/topics/insights/2020/spending-2020-together-on-twitter

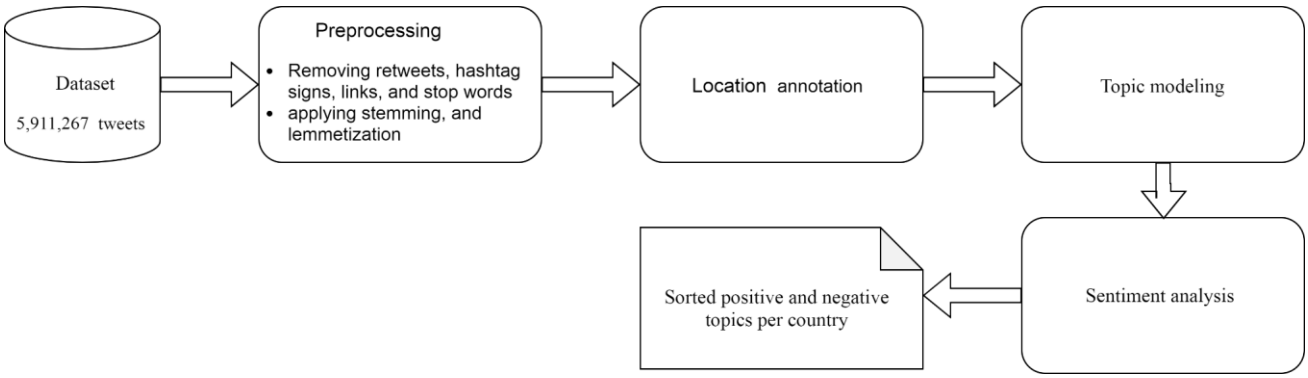


Fig. 1. The overall schema of the proposed method.

at home". This article uses more than five million English-written tweets that include the hashtags related to COVID-19 (Sarscov2, Corona, Coronavirus, Covid, COVID-19, and pandemic). These tweets were published from March 23 to June 23, 2020, at the same time as the start of global quarantines in most countries. The aim is to analyze the discussed topics in the first months of the pandemic.

A. *Preprocessing*

The preprocessing of textual data has a fundamental effect on the improvement of topic modeling [29]. High-level processing of raw tweets is difficult because they are too vague and unstructured. For preprocessing the tweets, the incoming tweets are divided into sentence structures and then words (tokens). The stop words, which are very frequent in the texts, are then removed. These words include prepositions, conjunctions, and so on. Symbols such as parentheses, commas, etc. are also removed in the process. In addition, the hashtag, user ID, and links are removed.

The next preprocessing step is stemming. Stemming is one of the most important steps in preprocessing, which has a great impact on improving the text mining process. Words appear in different ways depending on their semantic and grammar roles in the sentence. These different appearances indicate the different meanings of these words, but since they are all derived from the same stem, they are relatively close in meaning. Therefore, in many applications of natural language processing and information retrieval, it is necessary to convert all derivatives of a word into its stem, which is the simple form of the word.

The next preprocessing step is lemmatization, which is similar to stemming, except that it uses a dictionary of words to map all related words to the original form. In this paper, SnowballStemmer[6] and WordNetLemmatizer[7] algorithms are used for stemming and lemmatization, respectively. One of the advantages of the SnowballStemmer framework is its low error rate [30]. Finally, to increase the information retrieval rate, frequent terms in tweets such as COVID-19, Covid, Coronavirus, pandemic, and Trump are removed.

B. *Location extraction*

User data annotation is an important part of text processing. Common methods of location identification for each tweet are the geographical location registered in the user account or the current geographical coordinates of the user. Because the location of the content of a tweet is sometimes different from the user location, we analyze the content to specify the location. For example, if users from the United States, India, etc. share tweets about China, they receive a Chinese location tag, as this tweet has a discussion related to China. For this purpose, a comprehensive location lexicon containing 6756 words from the names of places related to 32 countries has been collected [9]. The investigated countries have the most number of COVID-19 affirmed cases and the most noteworthy number of tweets. They include the United States, United Kingdom, United Arab Emirates, Turkey, Switzerland, Sweden, Spain, South Korea, Singapore, Saudi Arabia, Russia, Qatar, Portugal, Peru, Pakistan, Netherlands, Mexico, Japan, Italy, Ireland, Iran, India, Germany, France, Ecuador, China, Chile, Canada, Brazil, Belgium, and Belarus. Location words have been collected using the GeoNames geographic database. They include country names, province names, state names, city names, and related titles. For China, for example, 3415 related place titles have been collected. Then, with the help of a text processing algorithm, the vocabularies of this dictionary are compared with tweets. Then, by using the normalization rules in the proposed algorithm, all the places related to a country are mapped to the country's name.

C. *Topic Modeling*

Topic modeling aims to find the main topics discussed in a collection of documents or corpus. Topic modeling algorithms such as Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) divide texts into several clusters so that the texts within each cluster are considered close together in terms of clustering criteria (for example, phrases that often appear together). Finally, the naming of topics for clusters (guessing the relationship of a cluster phrases) is manually performed by experts.

[6] http://snowball.tartarus.org/algorithms/porter/stem_ISO_8859_1.sbl

[7] http://search.cpan.org/~tpederse/WordNet-Similarity-2.05/lib/WordNet/stem.pm

We use LDA topic modeling algorithm, which assumes that documents consist of words, each belonging to a topic, and a document partially relates to several subjects. In this regard, a lexicon of words is considered so that each topic is a possible distribution on this set of words. Words are generated from the existing corpus for each document in two steps. In the first step, a possible distribution of the topics is selected randomly. In the second step, for each document word, a topic is chosen randomly using the first stage probability distribution. A word is then randomly selected with the specified probability distribution. In this regard, P ($w_{d,i}$) is the probability of the i^{th} word in document d [11]

$$P(w_{d,i}) = \sum_{j=1}^{V} P(w_{d,i}|z_{d,i} = Z_j) \times P(z_{d,i} = Z_j) \quad (1)$$

V is the total number of topics. $z_{d,i}$ is the topic assigned to $w_{d,i}$ and $z_{d,i} = Z_j$ denotes for the word $w_{d,i}$ which is assigned to the topic j [11].

To identify hot tweet topics using LDA, a word/text matrix should be first be created. In other words, the tweet dataset is expressed as a matrix. For this purpose, the keywords are extracted from the tweets by extracting the feature vector containing the weight of the tokens using the conventional Term Frequency - Inverse Document Frequency (TF-IDF) [31] method. Keywords are a set of important words in a text that describe the content of the text and can be used for various purposes, such as topic modeling. In estimating TF-IDF, a single term with more iterations in the document is more important and subsequently weighs more. The longer the document, the more likely to have a higher TF. For this reason, the TF should be normalized to the length of the document:

$$W_{t,d} = (1 + \log TF_{t,d}).\log \frac{N}{DF_t} \quad (2)$$

where d is a document, t is a term, $TF_{t,d}$ is the number of occurrences of t in d, DF_t is the number of documents in the set in which t occurs, N is the total number of documents in the corpus, and $W_{t,d}$ is the weight of TF-IDF for each term t in each document d [32]. In the case of inverse document frequency (IDF), the less a term is repeated in documents, the more valuable it is:

$$\mathrm{IDF(t)} = \log\left(\frac{N}{DF_t}\right) \quad (3)$$

According to Equation. 3, DF_t refers to the number of documents in which the term t is repeated. Finally, the weighted matrix of the words/documents is calculated. The LDA model is then used, which indicates the probability that a word is important in a text. By applying this model to the mentioned matrix, the obtained probabilistic values are considered as feature vectors. The probabilistic values for each word indicate the similarity of the meaning of that word to the topic in question. Finally, based on the obtained feature vector, a proper distinction can be made between the topics. Fig 2 shows the levels of an LDA algorithm.

In order to get better and more accurate results for LDA model training, retweets have been removed and 150,000 tweets have been randomly assigned to each week so that the weekly comparison of tweets is possible. Finally, the proposed model is implemented for 14 weeks. Table 1 shows the titles of the ten topics obtained along with the most relevant words of each topic. Which includes the names of the topics, the most relevant words of each topic with the possibility that the word belongs to that topic, and an example of user tweets.

In Fig.2, M refers to the number of documents and N indicates the number of words in the document [11]. The first level, or the level of the corpus, contains the parameters α and β, which are considered as examples for the corpus production process and their value is obtained experimentally. α represents the relative strength between the implicit topics in the corpus and β represents the probability distribution for the

TABLE I. TOPICS WITH MOST RELEVANT WORDS, THE PROBABI

Topics	The most relevant words	
Telecommuting	0.02 * "work" + 0.014 * "home" + 0.01 * "lockdown" + 0.007 * "stand"	
Death cases	0,033 * "case" + 0,018 * "death" + 0,015 * "report" + 0,014 * "total" + 0,011 * "number"	
Protests	0.014 * "protest" + 0.011 * "racist" + 0.01 * "black" + 0.01 * "leader" + 0.009 * "right"	
Second wave	0,02 * "second" + 0,012 * "surge" + 0,011 * "warn" + 0,011 * "prepare"	#equity markets are about to hit as well.
Reopening	0,011 * "reopen" + 0,009 * "safe" + 0,008 * "rule" + 0,008 * "increase" + 0,008 * "recovery"	" Gyms should reopen tf. I'll clock in and work during the corona crisis cus' i miss my workouts."
Anger expression	0,019 * "rally" + 0,009 * "stupid" + 0,007 * "sick"	" i'm sorry.. but y'all are so damn stupid if you're out still PARTYING with this corona virus going around."
Masking	0,019 * "mask" + 0,012 * "hand" + 0,007 * "face"	" Kindly use face masks to protect yourself and let's all fight against coronavirus #facemask."
Peak of the disease	0.014 * "spike" + 0.008 * "scientist"	"#RESIST "Corona virus is just beginning to spike in US. "
Medication	0.011 * "drug" + 0.007 * "dexamethason" + 0.007 * "save"	" Dexamethasone Is Proven to be the 1st COVID-19 cure.... It's not effective in early stages but very"
Social distance	0.032 * "social" + 0.032 * "distance" + 0.007 * "keep"	" Keep in mind that social distancing will reduce the need for hospital beds for all other infectious diseases."

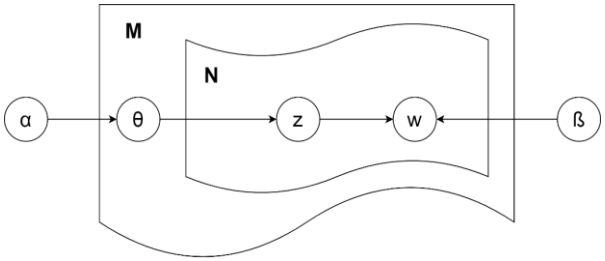


Fig.2. LDA graphical model..

implicit topic itself [11]. The second level, or document level, contains the parameter θ which is a one-time sample of each document and its value is obtained by the LDA algorithm. The third level, or word level, contains the parameters z and w, which are sampled once for each word of each document [11].

D. Sentiment Analysis

Analyzing users' sentiments about various phenomena is very important to identify positive and negative comments. In this article, text2data[8] user interface programming is used for sentiment analysis because it includes specially prepared classification models for Twitter trained on billions of manually verified entries and uses deep learning-based NLP methods. This programming interface uses cloud computing with automatic scale configuration. Every term is being split into chunks and represented as a tree structure. Sentiment analysis is applied to each cluster of topics, and it is possible to display the sentiment tree for each cluster. Fig.3 provides an example of a sentiment tree for the most relevant words on the topic of "protests." The words of this topic are shown in the tree's leaves, which are distinguished by different colors. Finally, the color of the root node reflects the sentiment of the whole tree. Table 2 shows the sentiment scores for the most relevant words in each topic, which contains the names of the topics and the sentiment scores. Sentiment score is a number between -1 to +1, which shows the most negative and positive sentiments, respectively. According to this table, the topics are divided into two categories including positive and negative topics as follows:

- Positive topics: "Reopening", "Telecommuting", "Masking", "Medicine", and "Social distance".
- Negative topics: "Death cases", "Protests", "Anger expression", "Second Wave", and "Peak of the disease".

IV. Experiment

The proposed method is implemented using the Python programming environment. The GATE Natural Language Processing Framework is used to tag the location of tweets related to each country. To preview tweets, the required functions have been called from the Genism library, which is one of the foremost prevalent libraries utilized in topic modeling.

By trial and error, we finally selected ten topics, as this number results in better diversity and less overlap. The results of the experiment are presented in four subsections. These subsections discuss the static analysis of sentiments for each country, the importance of each topic among countries, the top topics of the countries, and finally, the trends of the negative and positive topics for the countries.

A. Static analysis of sentiments

This subsection provides the percentage of positive and negative tweets for the mentioned countries. Fig.4 shows the percentage of positive and negative tweets for each country. Among the total tweets, 53.60% of the tweets are negative and 46.39% are positive. We assume countries with about more than 60% positive and negative tweets as positive countries and negative countries, respectively. Belarus ranks first in negative countries and Saudi Arabia ranks first in positive countries. The list of these countries is as follows:

- Negative countries: Belarus, Peru, Portugal, Brazil, South Korea, Belgium, Sweden, China, Russia, and Canada.
- Positive countries: Saudi Arabia, Ecuador, Iran, Qatar, Singapore, France, and Germany.

B. Importance of each topic

The importance of each topic varies from country to country. Fig.5 shows the share of each topic among the total tweets of each nation. The charts of positive and negative topics are shown in green and red, respectively. The top four countries that have addressed the positive topics are:

Reopening: Saudi Arabia, Iran, Turkey, and Qatar.

Telecommuting: India, USA, Australia, and Japan.

Social distance: USA, India, UK, and China.

Medicine: Mexico, Singapore, Ecuador, and Chile.

Masking: Ecuador, Germany, France, and the UK.

On the other hand, the top four countries that have addressed negative topics are as follows:

- Death cases: Belarus, Peru, Portugal, and Sweden.
- Peak of the disease: Russia, South Korea, Portugal, and Italy.
- Protests: Belgium, China, the Netherlands, and the United Arab Emirates.
- Second wave: Canada, Spain, Sweden, and Brazil.
- Anger expression: Ireland, Brazil, Mexico, and Turkey.

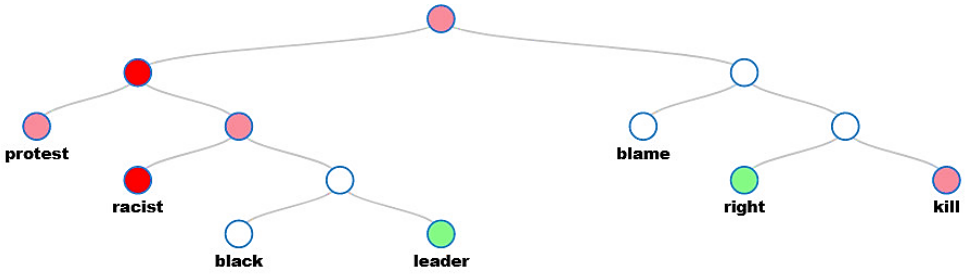


Fig.3. Sentiment tree visualization of the most relevant words for topic "protest". Green, white, light red, and red nodes represent positive, neutral, negative and very negative sentiments, respectively.

[8] https://text2data.com

TABLE II. SENTIMENT SCORE BETWEEN -1 AND +1 FOR EACH.

Topic	Sentiment score
Telecommuting	+0.554
Death cases	-0.547
Protests	-0.569
Second wave	-0.493
Reopening	+0.582
Anger expression	-1.000
Masking	+0.161
Peak of the disease	-0.131
Medication	+0.186
Social distance	+0.640

C. *Top Topics of Countries*

Table 3 illustrates the top four topics for these 32 countries. The topics of death cases and reopening have been the major issues in most countries. Death cases and peak of the disease are the hottest negative topics between countries. On the other hand, reopening is the hottest positive topic between countries.

Amongst other topics, Canada, Turkey, and Brazil have addressed the topics of the second wave and anger expression. In Belgium, China, Iran, and India, the topics of the peak of the disease and protests are important. Users have expressed their anger about the failure to close borders in Iran and non-compliance with social distancing within the early stages of the virus flare-up in India. In Ecuador and UK, the topics of medicine and masking have been important. British researchers' discovery of the effect of dexamethasone on patients' recovery for the first time made this topic more important in this country. In the USA, Singapore, and Russia, the topics of the peak of the disease and medicine have been important. According to US statistics, it had the highest incidence of the disease among other countries, and the importance of the peak of the disease topic in this country is evident.

D. *Trend of Negative and Positive Topics*

The cumulative process of positive versus negative topics can provide useful information. In this regard, the trends of frequency of tweets related to negative and positive topics are presented in Fig.6 for three months. For better analysis, the frequency trend of the number of confirmed COVID-19 cases as reported by the World Health Organization is also provided in black. The two vertical axes show the number of tweets and the number of confirmed cases in 14 weeks. It can be seen that the trend of negative topics has a high correlation with the number of confirmed cases for many countries. This fact also validates the proposed method.

E. *Comparison*

As the subject of this article is the analysis of corona-related tweets including location extraction, sentiment analysis, and topic trend analysis, a comparison with related research is provided below. Table 4 summarizes the comparison results between the proposed method and previous research in terms of sentiment analysis, location analysis, and topics trends analysis.

Xue et al. [13], Abd-Alrazaq et al. [25], Chandrasekaran et al. [26], and Boon-Itt and Skunkan [27], have not performed any regional analysis, and Klaifer et al. [28], Doogan et al. [8], Jang et al. [10] have focused on a limited number of countries. The proposed study is conducted for the 32 countries most affected by the coronavirus. The top topics related to each country are presented and the share of discussion on each topic is specified. Also, the cumulative trends of negative as well as positive topics for these countries have been analyzed.

V. CONCLUSION

In this article, the top topics of user tweets in 32 countries more involved with COVID-19 have been identified. The LDA algorithm has been used to detect these topics and the top ten topics were obtained. Then a lexicon-based method has been proposed to extract the location content of each tweet. Moreover, sentiment analysis was performed on the most relevant words of each topic, and the topics were divided into two groups including positive and negative. The positive group includes the topics of reopening, masking, telecommuting, medicine, and social distancing. The negative group includes the topics of death cases, peak of the disease, protests, second wave, and anger expression. In general, negative topics make up a larger percentage of countries' tweets. European nations such as the UK, Germany, and France have discussed more on the masking topic. Eastern countries such as India, China, and Singapore have paid more attention to the death cases topic. Tweets from USA also paid more attention to the topic of death cases, which could be due to the country's greater conflict and the high number of patients in the early stages of the epidemic. The time trends of positive and negative topics of countries have shown that there has been a high correlation between the amount of discussion on negative topics and the number of confirmed cases.

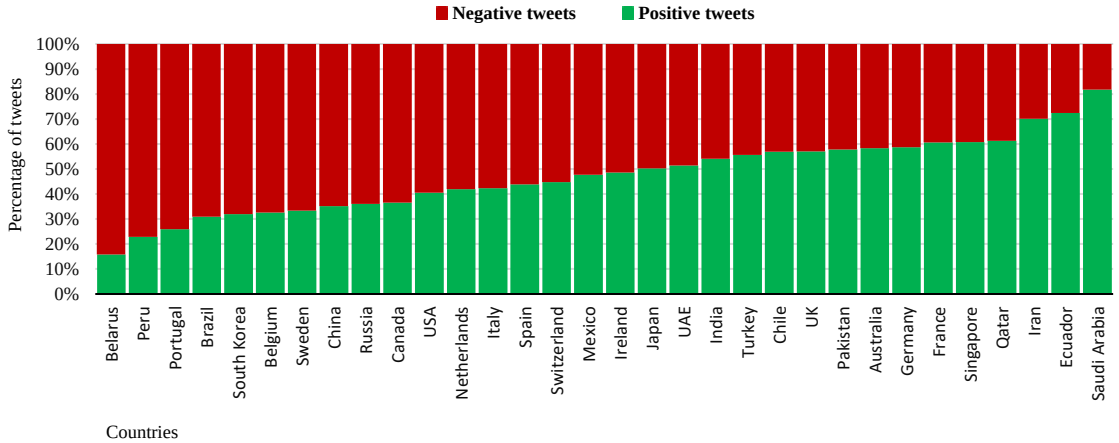


Fig.4. Percentage of negative and positive tweets for each country.

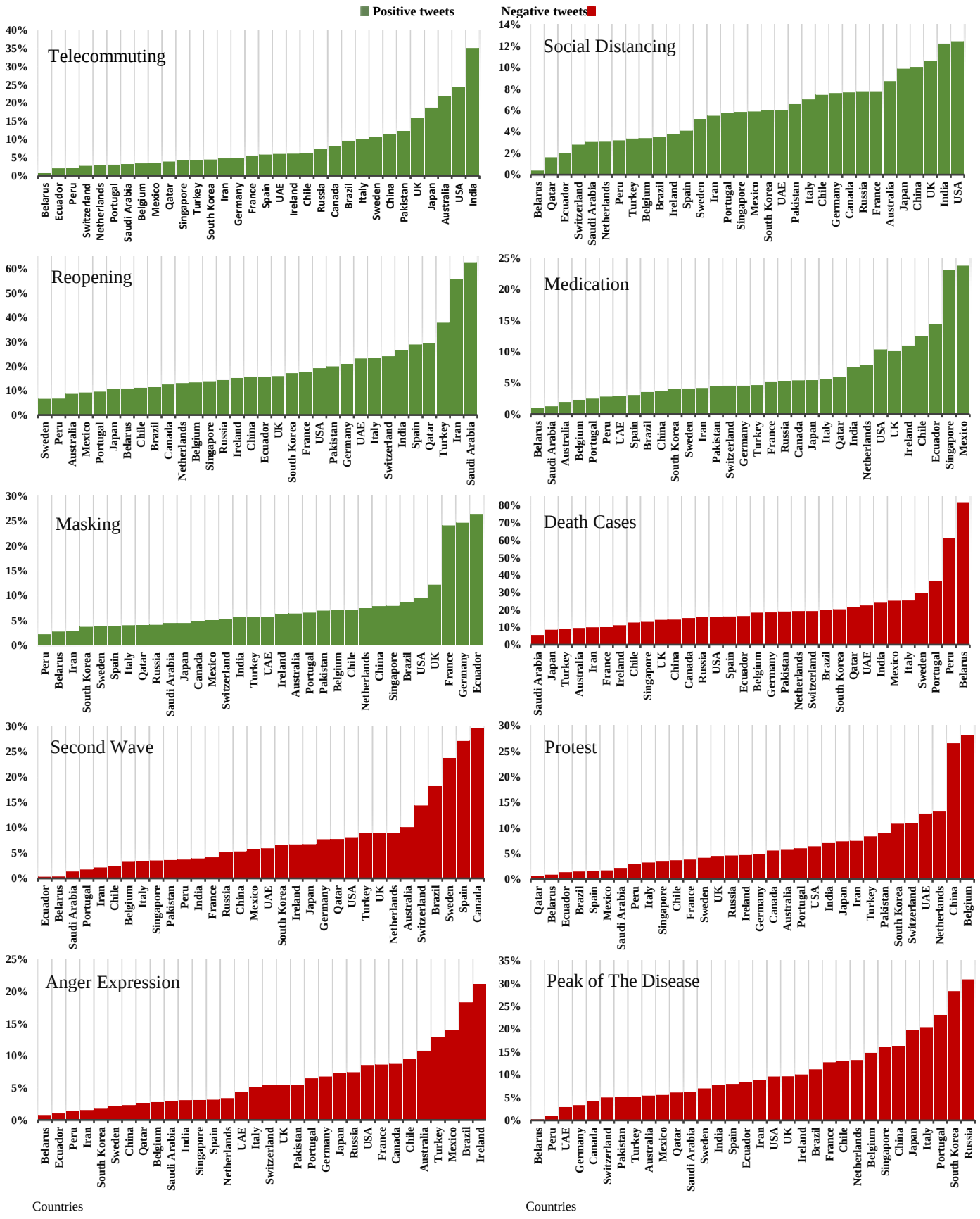
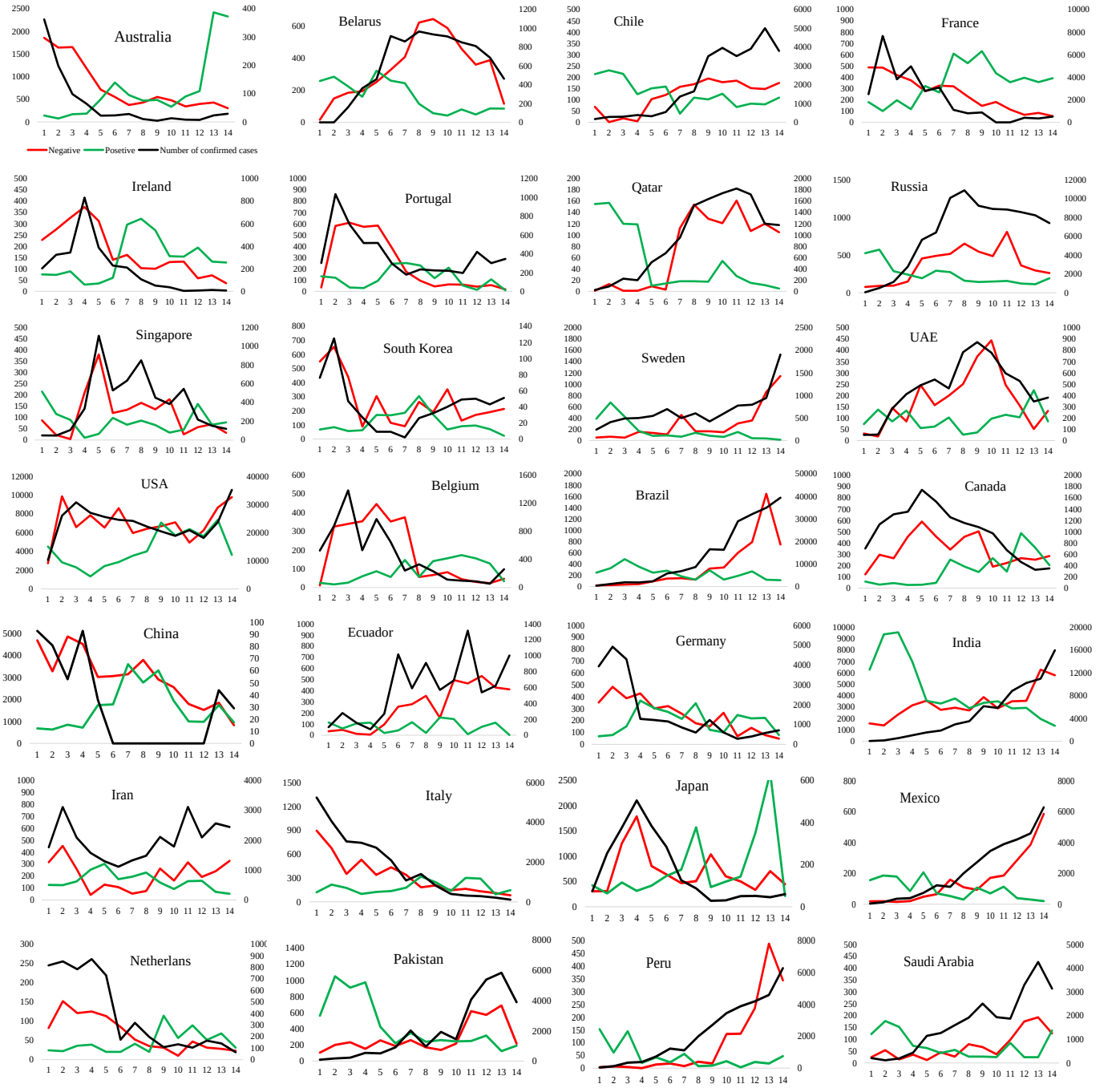


Fig 5 The percentage of tweets for each topic per country

TABLE III. IMPORTANCE OF FOUR TOP TOPICS AMONG 32 COUNTRIES INVOLVED MORE WITH COVID-19. ■ FIRST DISCUSSED TOPIC. ■ SECOND DISCUSSED TOPIC. ■ THIRD DISCUSSED TOPIC. ■ FOURTH DISCUSSED TOPIC.

Topic \ Country	Australia	Belarus	Belgium	Brazil	Canada	Chile	China	Ecuador	France	Germany	India	Iran	Ireland	Italy	Japan	Mexico	Netherlands	Pakistan	Peru	Portugal	Qatar	Russia	Saudi Arabia	Singapore	South Korea	Spain	Sweden	Switzerland	Turkey	UAE	UK	USA
Death cases																																
Reopening																																
Peak of the disease																																
Second wave																																
Telecommuting																																
Anger expression																																
Protests																																
Masking																																
Medication																																
Social distance																																



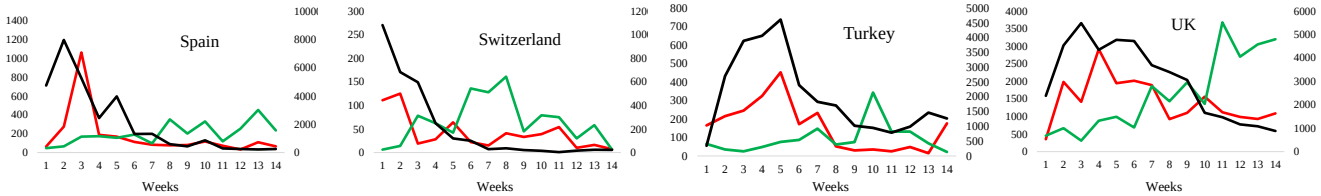


Fig.6. The cumulative trend of positive and negative topics.

TABLE IV. A COMPARISON BETWEEN THE PROPOSED METHOD AND PREVIOUS RESEARCHES BASED ON TOPIC MODELING ON COVID-19 TWEETS.

Reference	Sentiment	Location	Topic trends
Xue et al [13]	✓	×	×
Abd-Alrazaq et al [25]	✓	×	×
Chandrasekaran et al [26]	✓	×	×
Boon-Itt and Skunkan [27]	✓	×	×
Klaifer et al [28]	✓	Brazil and USA	×
Doogan et al [8]	×	Australia, Canada, New Zealand, Ireland, UK, and USA	×
Jang et al [10]	✓	Canada and USA	✓
Azizi et al [9]	×	USA, China, UK, and Canada	✓
Proposed methodology	✓	USA, UK, UAE, Turkey, Switzerland, Sweden, Spain, South Korea, Singapore, Saudi Arabia, Russia, Qatar, Portugal, Peru, Pakistan, Netherlands, Mexico, Japan, Italy, Ireland, Iran, India, Germany, France, Ecuador, China, Chile, Canada, Brazil, Belgium, Belarus, and Australia	✓

Although this study examined a large number of users' tweets, the Twitter social network cannot reflect the thoughts and opinions of all people in a community. Also, the use of non-English language tweets cannot precisely represent the public opinions in non-English speaking countries. In the future, people's perspectives on specific areas such as psychology, politics, economics, and education can be evaluated by extending this work.

REFERENCES

[1] D. Pruss et al., "Zika discourse in the Americas: A multilingual topic analysis of Twitter," PloS one, vol. 14, p. e0216922, 2019.

[2] M. Cai, J. Li, M. Nali, and T. K. Mackey, "Evaluation of Hybrid Unsupervised and Supervised Machine Learning Approach to Detect Self-Reporting of COVID-19 Symptoms on Twitter," in International Conference on Communications Workshops Montreal, 2021: IEEE, pp. 1-6.

[3] T. Mackey et al., "Machine learning to detect self-reporting of symptoms, testing access, and recovery associated with COVID-19 on Twitter: retrospective big data infoveillance study," JMIR public health and surveillance, vol. 6, p. e19509, 2020.

[4] M. Gupta, A. Bansal, B. Jain, J. Rochelle, A. Oak, and M. S. Jalali, "Whether the weather will help us weather the COVID-19 pandemic: Using machine learning to measure twitter users' perceptions," International journal of medical informatics, vol. 145, p. 104340, 2021.

[5] F. Salmani, H. Vahdat-Nejad, and H. Hajiabadi, "Analyzing the Impact of COVID-19 on Economy from the Perspective of Users Reviews," presented at the Eleventh International Conference on Computer and Knowledge Engineering, Mashhad, 2021.

[6] D. M. Blei and J. D. Lafferty, "Topic models," Text mining: classification, clustering, and applications, vol. 10, p. 34, 2009.

[7] I. Kagashe, Z. Yan, and I. Suheryani, "Enhancing seasonal influenza surveillance: topic analysis of widely used medicinal drugs using Twitter data," Journal of medical Internet research, vol. 19, p. e315, 2017.

[8] C. Doogan, W. Buntine, H. Linger, and S. Brunt, "Public perceptions and attitudes toward COVID-19 nonpharmaceutical interventions across six countries: a topic modeling analysis of twitter data," Journal of medical Internet research, vol. 22, p. e21419, 2020.

[9] F. Azizi, H. Vahdat-Nejad, H. Hajiabadi, and M. H. Khosravi, "Extracting Major Topics of COVID-19 Related Tweets," presented at the Eleventh International Conference on Computer and Knowledge Engineering, Mashhad, 2021.

[10] H. Jang, E. Rempel, D. Roth, G. Carenini, and N. Z. Janjua, "Tracking COVID-19 discourse on twitter in North America: Infodemiology study using topic modeling and aspect-based sentiment analysis," Journal of medical Internet research, vol. 23, p. e25431, 2021.

[11] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," Journal of machine Learning research, vol. 3, pp. 993-1022, 2003.

[12] H. Cunningham, "GATE, a general architecture for text engineering," Computers and the Humanities, vol. 36, pp. 223-254, 2002.

[13] J. Xue, J. Chen, C. Chen, C. Zheng, S. Li, and T. Zhu, "Public discourse and sentiment during the COVID 19 pandemic: Using Latent Dirichlet Allocation for topic modeling on Twitter," PloS one, vol. 15, p. e0239441, 2020.

[14] I. Kagashe, Z. Yan, and I. Suheryani, "Enhancing seasonal influenza surveillance: topic analysis of widely used medicinal drugs using Twitter data," Journal of medical Internet research, vol. 19, no. 9, p. e315, 2017.

[15] D. Pruss et al., "Zika discourse in the Americas: A multilingual topic analysis of Twitter," PloS one, vol. 14, no. 5, p. e0216922, 2019.

[16] M. D. T. Nzali, S. Bringay, C. Lavergne, C. Mollevi, and T. Opitz, "What patients can tell us: topic analysis for social media on breast cancer," JMIR medical informatics, vol. 5, p. e23, 2017.

[17] A. Garcia-Rudolph, S. Laxe, J. Saurí, and M. B. Guitart, "Stroke survivors on twitter: sentiment and topic analysis from a gender perspective," Journal of medical Internet research, vol. 21, p. e14077, 2019.

[18] J. X. Koh and T. M. Liew, "How loneliness is talked about in social media during COVID-19 pandemic: text mining of 4,492 Twitter feeds," Journal of Psychiatric Research, vol. in press, 2020.

[19] J. Xue, J. Chen, C. Chen, R. Hu, and T. Zhu, "The Hidden Pandemic of Family Violence During COVID-19: Unsupervised Learning of Tweets," Journal of medical Internet research, vol. 22, p. e24361, 2020.

[20] J. Yu, Y. Lu, and J. Muñoz-Justicia, "Analyzing Spanish News Frames on Twitter during COVID-19—A Network Study of El País and El Mundo," International Journal of Environmental Research and Public Health, vol. 17, p. 5414, 2020.

[21] E. Fino, B. Hanna-Khalil, and M. D. Griffiths, "Exploring the public's perception of gambling addiction on Twitter during the COVID-19

pandemic: Topic modelling and sentiment analysis," Journal of Addictive Diseases, vol. 39, pp. 1-19, 2021.

[22] D. Valle-Cruz, V. Fernandez-Cortez, A. López-Chau, and R. Sandoval-Almazán, "Does Twitter Affect Stock Market Decisions? Financial Sentiment Analysis During Pandemics: A Comparative Study of the H1N1 and the COVID-19 Periods," Cognitive computation, vol. 13, pp. 1-16, 2021.

[23] J. Hacker, J. vom Brocke, J. Handali, M. Otto, and J. Schneider, "Virtually in this together–how web-conferencing systems enabled a new virtual togetherness during the COVID-19 crisis," European Journal of Information Systems, vol. 29, pp. 1-22, 2020.

[24] J. C. Lyu, E. Le Han, and G. K. Luli, "COVID-19 Vaccine–Related Discussion on Twitter: Topic Modeling and Sentiment Analysis," Journal of Medical Internet Research, vol. 23, p. e24435, 2021.

[25] A. Abd-Alrazaq, D. Alhuwail, M. Househ, M. Hamdi, and Z. Shah, "Top concerns of tweeters during the COVID-19 pandemic: infoveillance study," Journal of medical Internet research, vol. 22, p. e19016, 2020.

[26] R. Chandrasekaran, V. Mehta, T. Valkunde, and E. Moustakas, "Topics, trends, and sentiments of tweets about the COVID-19 pandemic: Temporal infoveillance study," Journal of medical Internet research, vol. 22, p. e22624, 2020.

[27] S. Boon-Itt and Y. Skunkan, "Public perception of the COVID-19 pandemic on Twitter: Sentiment analysis and topic modeling study," JMIR Public Health and Surveillance, vol. 6, p. e21978, 2020.

[28] K. Garcia and L. Berton, "Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA," Applied Soft Computing, vol. 101, p. 107057, 2021.

[29] M. J. Denny and A. Spirling, "Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it," Political Analysis, vol. 26, pp. 168-189, 2018.

[30] A. G. Jivani, "A comparative study of stemming algorithms," Int. J. Comp. Tech. Appl, vol. 2, pp. 1930-1938, 2011.

[31] J. Ramos, "Using tf-idf to determine word relevance in document queries," in First instructional conference on machine learning, Washington, 2003: Citeseer, pp. 29-48.

[32] K. Church and W. Gale, "Inverse document frequency (idf): A measure of deviations from poisson," in Natural language processing using very large corpora: Springer, 1999, pp. 283-295.