Social Media and Epidemic Modeling (Corona)

Team 3

ZHANG, Cao
ID: 20803018
czhangdf@connect.ust.hk

LU, Zijun
ID: 20784286
zlubc@connect.ust.hk

HUANG, Xi
ID: 20785694
xhuangcg@connect.ust.hk

Abstract

The coronavirus pandemic has infected over 80 million people worldwide since it was first known. This project aims to analyze people's real concerns through the Twitter media during the COVID-19, and explore three research questions *"Which region is the most engaged in debating the epidemic on Twitter?"*, *"How do people's attitudes to the epidemic change over time in the different regions?"* and *"What topics are most concerned about when discussing the epidemic in different regions?"* with their critical analysis and findings. The purpose of this paper was to help the government's decision if the pandemic occurs again in the future.

ACM Reference Format:
ZHANG, Cao, LU, Zijun, and HUANG, Xi. 2022. Social Media and Epidemic Modeling (Corona): Team 3. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

In 2020, WHO declared Covid-19 as a global pandemic, and until this movement, Covid-19 has affected human society for two years. The development of the worldwide pandemic brought many challenges to human society, not only in the medical domain but also in the economic, social, and political domains. The governments of many countries take action to stop the spread of disease, including lockdown, quarantine, and promotion of vaccines. Due to the rise of social media in this era, Twitter has become one of the most extensively used social media worldwide. And it is common to know users' viewpoints by analyzing their tweets because the text found on Twitter seems to be a good proxy for public perception about the current pandemic [1].

Social media has both advantages and disadvantages in helping society overcome the pandemic [2]. The spreading of misinformation on social media can heavily influence people's behavior, affecting the effectiveness of response measures in disaster management. For example, the anti-vaccination message on social media decreases vaccine hesitancy [3]. Therefore, detecting and characterizing people's discussion topics on social media is vital for future analysis, such as anticipating rumors and studying the propagation of social media topics.

Significantly, the topics discussed on Twitter regarding the pandemic in the different regions reveal the aspects of the crisis that are considered more important and salient for the population of a particular area. Also, the sentiment polarity of tweets can provide valuable information about government measures, for example, social distancing and travel bans. Knowing the changes in sentiment polarity through time and interpreting those changes with the major events and government decisions may help sociality predict how similar measures will affect the population.

2 Related Work

The information encoded in the text contexts created by Twitter users provides valuable indications and can be helpful for researchers and experts to extract meaningful information. [4] for instance, by using the Reddit data to monitor the change in public concerns and attitudes towards the COVID-19 pandemic. In particular, their study shows the utility of Latent Dirichlet Allocation (LDA) topic modeling in discovering the public's concern, which is strongly related to our project. Our study uses the LAD model to analyze Twitter plain-text data and abstract topics from the text. In addition, the data preprocessing and high-frequency word analysis method mentioned in this paper will also be one of the references.

Regarding the public sentiment analysis, the study conducted in research [5] explored the public perception of the COVID-19 at its early stage in China. They aimed to extract and identify public opinion from the social media, Weibo, and analyze the spatial-temporal characteristics based on Weibo texts. Similarly, they used the LDA model and RF algorithm to identify the topic of the text. Time and spatial analysis based on the statistical solution. We used the statistical solution to conduct the geographical research for our study. More importantly, this paper shows us the sentiment analysis based on implementing machine learning models such as SVM or VADER to identify the public's sentiments rather than extract sub-topics.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

The geolocation information is essential to discovering the discussion engagement towards the Covid-19 pandemic on Twitter. [6] This study aimed to analyze the sentiment variation trend before and after the announcement of Pfizer's vaccine and the difference in people's sentiment changes according to different geographic locations. They used a dictionary to map the location information in the user profile for each tweet to the country level to acquire the location information. The study performed the geographic analysis at the national and state levels. Likewise, our study uses a dictionary to map the self-defined location information to the country level.

3 Problem Definition (Research Questions)

By knowing the viewpoints and attitudes of the public correctly and timely, the government and social media can disseminate and update epidemic-related information to avoid mass panic [7]. Besides, the government or policymakers can get feedback on policies and regulations they have issued before, such as the vaccine's launch, and make proper adjustments by knowing the public's attitude [8]. And they also may know which country's policy is worth learning from by comparing public sentiments of different regions.

Moreover, the public's attitude change reflects the epidemic trend; for example, if they become positive from negative, it may indicate that the condition is getting better and better.

In agreement with the motivation discussed above, the research questions addressed in this study can be summarized as follows:

RQ1: Which region is the most engaged in debating the epidemic on Twitter?
RQ2: How do people's attitudes to the epidemic change over time in the different regions?
RQ3: What topics are most concerned about when discussing the epidemic in different regions?

4 Data

4.1 COVID-19 Timeline

We organized a timeline of Covid-19 development, including some significant events related to the pandemic [9], as shown in Figure. 1.

The first known infections from COVID-19 were discovered in November 2019 in China. Later, the virus started to spread worldwide in January 2020. COVID-19 has caused about 503 million suspected infection cases and 6.19 million deaths globally [10].

Our project focuses on the evolution of the COVID-19, so we decide to divide the whole timeline into three stages:

(1) The pre-pandemic is from January 2020, when the virus started spreading worldwide, to November,2020.
(2) The mid-pandemic is from December 2020, when the Alpha variant became dominant, to October 2021.

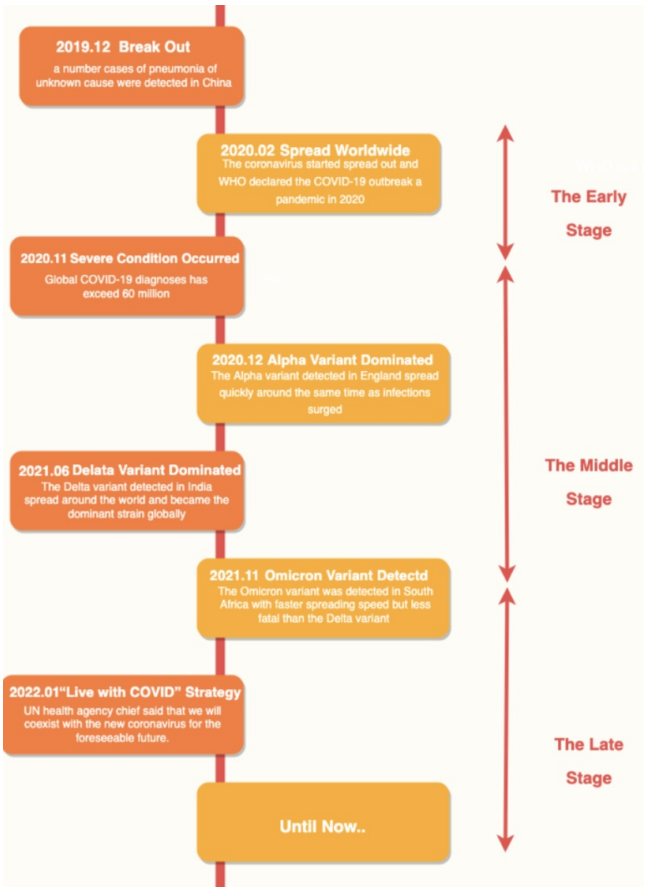


Figure 1. Timeline of the COVID-19

(3) The post-pandemic is from November 2021, when the Omicron variant was detected, until now.

4.2 Data Collection and Integration

We divided the dataset into three main parts. Each part contains tweets from pre-pandemic, mid-pandemic, and post-pandemic periods respectively, in line with the timeline outlined in the previous timeline.

The pre-pandemic part consists of two different datasets, one contains tweets from March to April 2020 [11], and another includes tweets from July to August 2020 [12]. Both datasets are sourced from Kaggle, a website providing numerous open-source datasets. The mid-pandemic datasets hold tweets from November to December 2020 [13] provided by this Social Network course at HKUST. The post-pandemic dataset is composed of 3 datasets, one containing tweets in December 2021 [14], one containing the tweets from December 2021 to January 2022 [15], and the last one containing tweets from February to March 2022 [16].

Since we acquired datasets from different sources, the attributes of each dataset vary from the other. For example, some dataset has the column "*user_url*," while the others do not. To minimize the missing data of columns, we only

retained the common columns for most of the dataset. For instance, only one or two datasets do not have the column "*user_verified*," while other datasets do, we still keep this column. Another challenge is that although most datasets may have the column "verified users," they are not uniformly named. We also kept the naming uniform. Further, we leveraged two python libraries, geonamescache, and pycountry, to map the city with the country because most datasets only contain information about cities instead of countries. The processed dataset contains 13 columns including "*country_code*", "*user_id*" and etc. We mainly used five of these columns: "*user_id*", "*user_name*", "*country_code*", "*tweet_created_at*", and "*tweet*" in this project.

Finally, we generated three folders to represent the three different stages. Each folder contains CSV files for several different countries and a JSON file to store the number of tweets for each country. At the same time, we put the number of tweets of each country before the file name of the corresponding country's CSV file. For example, "179064_*UnitedStates.csv*" indicates 179,064 tweets in the United States.

4.3 Data Preprocessing

A high-quality dataset is essential for our further analysis. There is unreliable and valueless data in our datasets, which affect the effectiveness of our analysis, so we used some tools to preprocess the dataset.

As Twitter users come from different countries, tweets may contain multiple languages. Our study focused on analyzing English tweets, so we filtered out duplicate and non-English tweets by lang id [17], a standalone language identification tool. We further removed irrelevant information, such as URL links.

The required input format is different for the two models used in our study, so the further preprocessing results also are different. The VADER model pays more attention to the emotions of tweets, and information such as punctuation marks or emojis will affect the emotions of the sentence so that we retain the entire tweet. The LDA model is more focused on the potential concern of users, so information such as punctuation, emoji, numbers, retweet tag, and stop words are discarded to preserve the central theme better.

5 Methodology

5.1 Geographic analysis (RQ1)

The user's location information is self-defined on Twitter. Therefore, the data format is not uniform, and it may contain some useless data, other than location information. For example, in some tweets, users use city names as location information. To map from the city name to the country name, we used Geonames Cache, a python library, to search for country names based on the city name. And some location information is represented in the country code. To uniform the information type, we use the pycountry library to map from country code to country name.

We performed the geographic analysis on the national level for each country (a total of 219 counties), and we conducted the statistical analysis of the tweets number. As a result, we formed the tweets number distribution worldwide.

5.2 VADER for attitude analysis (RQ2)

Presented by C.J. Hutto and Eric Gilbert [18], Valence Aware Dictionary for sEntiment Reasoning(VADER) is a rule-based model for sentiment analysis. Compared with other machine-learning-based methods such as SVM [19], there is no need to use the previously seen texts to determine the sentiments of new texts for VADER since it is based on a gold-standard sentiment lexicon, which is suitable for microblog-like texts like tweets. C.J. Hutto stated that the result presented by the VADER is evaluated well in the social media domain – the correlation coefficient shows that VADER (r = 0.881) performs as well as individual human raters (r = 0.888).

The working process of VADER, as shown in Figure. 2 aims to map words in the sentence to sentiment scores in its pre-built dictionary and will generate positive, negative, and neutral values for the whole sentence. There will be a compound score that mathematically combines the positive, negative, and neutral values. And we use this compound value to evaluate whether a sentence is positive, negative, or neutral based on the standard provided in VADER(The sentiment is positive when compound value >= 0.05, neutral when -0.05 < compound value <0.05, and negative when compound value <=-0.05).

Figure 2. VADER Workflow

We intended to use the VADER to analyze the sentiments expressed in tweets based on those factors:

1. The number of tweets data for each country is limited so it is hard to get the expected result by training model
2. The length of the tweet text without a context is short so it is better to use a rule-based model after balancing the efficiency and the accuracy
3. Tweet texts may contain lots of emojis which can be processed by VADER to improve the accuracy of sentiment results
4. More accurate compared with another lexicon method which will be discussed in details in Section 5.4.1.

We tried a sentiment test based on one dataset, including 30 tweets, to check whether it was practical. And the result is presented in the Figure. 3:

1. There are seventeen tweets labeled "positive," about 56.67%.
2. There are eight tweets labeled "negative," about 26.67%.
3. There are five tweets labeled "neutral," about 16.66%.

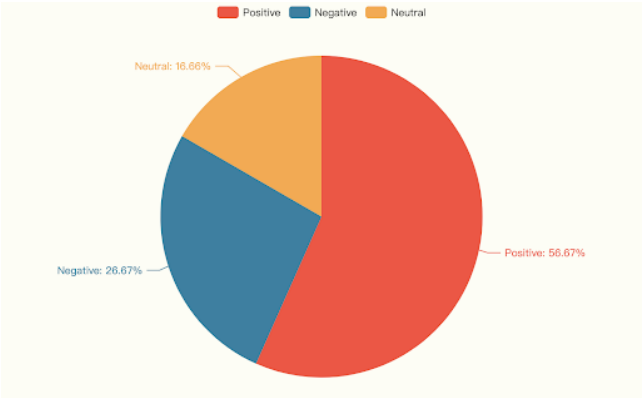


Figure 3. VADER Workflow

We compared the results with human-labeled results and found that around 73.33% of the result was the same as expected.

5.3 LDA for topic extraction (RQ3)

5.3.1 Latent Dirichlet Allocation model (LDA).

Originally formulated by Blei et al. [20], Latent Dirichlet Allocation (LDA) is an unsupervised learning model. The goal of LDA is to classify the target documents into different clusters, and then find short descriptions or the main topics of different clusters while preserving the essential statistical relationship, such as the similarity between collections. Blei argued in his paper that several models with similar functions, such as pLSA model, will suffer from a serious overfitting issue, that is the phenomenon of matching a particular data set too closely or precisely to fit other data well or predict future observations, while LDA model can easily avoid this problem.

After receiving the desired number of topics M and words N from the user, the LDA model can easily find the top-M representative topics for the document, and further find top-N keywords from the document for each topic. The Figure. 4 illustrates a SpongeBob document example result, where the M and N are both assumed as 3.

As mentioned earlier, our project aims to analyze the main topics of Twitter users in different countries during the COVID-19, which could reflect users' real cons. LDA is exactly what we expected, since each processed tweet can be represented as a target document. Also, Keras open-source code library will provide the visualization tool of the LDA model. Thus, the LDA model was considered as our topic extraction tool for solving the RQ3.

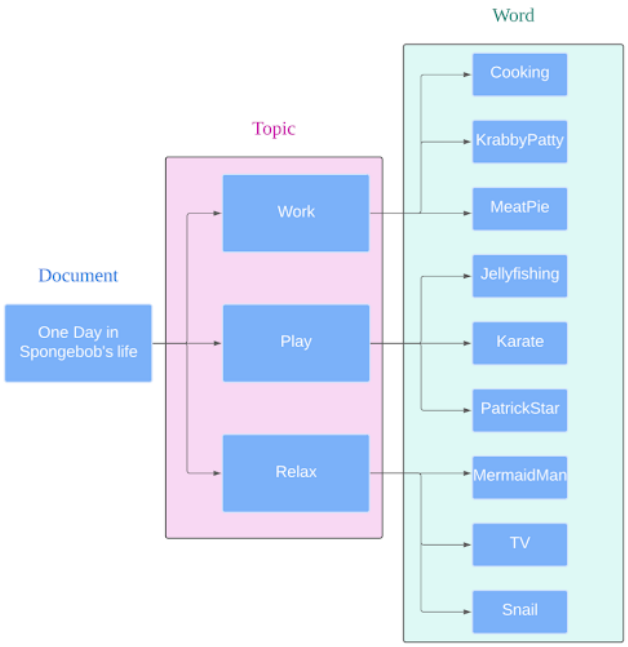


Figure 4. One Day in Spongebob's life

5.3.2 The workflow of LDA model.

In the LDA model, the target document is represented by a vector, which is also called a bag of words. The Figure. 5 shows the workflow of the LDA model. First, the user needs to give the target document and two priori parameters, α and β, which are just two integer numbers for generating two Dirichlet distributions. In the SpongeBob example mentioned in Section 6.4.1, the target document will be transformed into the LDA model to find a single topic from three candidate topics by the distribution generated by α. Similarly, the LDA model will utilize the chosen topic to find a single keyword from three candidate keywords by another distribution generated by β. This whole process only generates only one keyword. By iteratively processing this step, the LDA model finally generates a lot of keywords, which is a new document. However, the document, right now, is just some random keywords without any meaning. Thus, the LDA model needs to compare with the original document, to update three learnable parameters, until the LDA model can generate a document that is very close to the real document. Furthermore, two learnable distributions are what we want.

Given the expected number of topics, we can summarize the standard process into two steps:

(1) For each document, find a topic from topic distribution generated by a priori hyper-parameter α given by the user.
(2) For each topic founded on the previous step, find a word from word distribution generated by a priori hyper-parameter β given by the user.

We choose a subset from Armenia in the early stage to test the LDA model, where both M and N are 5. The Figure. 6 shows the visualization of topic 1 with the most relevant words.

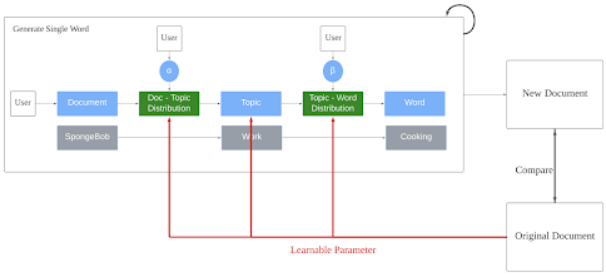


Figure 5. The workflow of LDA model

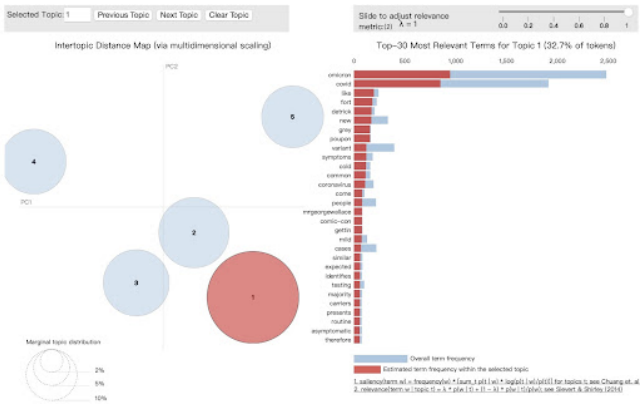


Figure 6. Visualization of tweets from Armenia

5.3.3 Salient Term in LDA model.

After multiple rounds of subset testing, we found that the LDA model could not adapt well to our project. The Figure. 7 demonstrates a poor result, that is, most topics overlap with each other. That is the main challenge we find in this project.

Other projects may collect all the tweets over a period of time, and those tweets cover a wide range of categories, such as politics and sport, which causes the LDA model to be easy to identify each topic. However, our dataset only contains the COVID-19 related tweets, and most tweets will contain the keyword "COVID" or "COVID-19", which is a leading cause of the high similarity score of some words.

Thus, the top 30 most salient terms were considered the primary measure metrics in this project to analyze users' concerns. The salient terms [21] are also produced by the LDA model during the distribution learning process. They can be thought of as a metric used to identify the most informative for identifying topics. The higher saliency value indicates that the word is more valuable.

This project splitted all the tweets into six categories based on salient terms. The Table. 1 shows several examples of each category. Other words, such as "UnemploymentRate", are equally crucial, but with few similar words, they are thus grouped into the "News Information" categories.

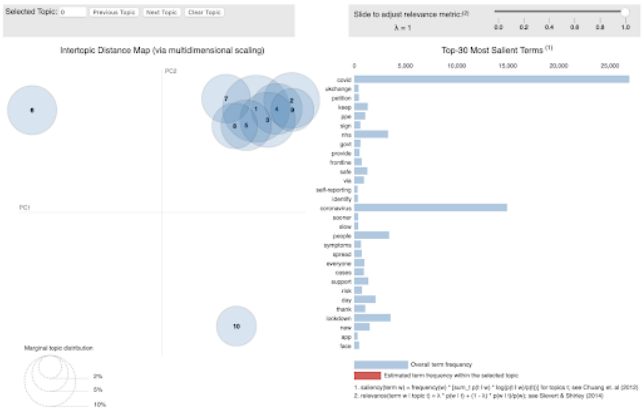


Figure 7. Poor result of LDA

5.4 Evaluation

We randomly sampled a subset of Luxembourg containing 105 tweets for our assessment. For fairness, all research questions share the same dataset. In addition, we adopted four measure metrics, and the formulas are as Table 2 and Table 3.

5.4.1 Evaluation of VADER.

Each tweet of Luxembourg was identified and labeled by each group member, and we recorded the final result in the form of a vote. Since there are only three members in our group, there would not be a tie. However, owing to different concepts, three members would produce three different results (1 vote for positive, neutral, and negative, respectively) for the same tweet on rare occasions. In this case, our group would discuss and vote again. We compared the results of manual judgment with those generated by two different models, and the results are shown in Table. 4.2 ˜ Table. 4.3 The VADER is more accurate in judging positive and neutral tweets, while TextBlob is better at judging negative tweets. In general, with higher accuracy, the VADER model was finally adopted in this project.

5.4.2 Evaluation of LDA.

Similar to Section 5.4.1, we artificially labeled each tweet into six categories, and compared the results with those generated by salient terms. The evaluations are shown in Table. 5.

6 Results

6.1 Geographic data summary (Tweets number distribution globally)

We mapped the tweets counts for each country using the accumulated COVID-related tweets and formed a heat map. The heat map as shown in Figure. 8 shows that certain regions produce numerous tweets, whereas others make a small amount. Initially, 219 countries were extracted from the dataset, and we narrowed down the sample countries to the top 10 countries with the most tweets. The filtered countries include the United States (379, 179 tweets), United Kingdom (143, 420 tweets), Thailand (73, 925 tweets), India (63, 165 tweets), Canada (31, 476 tweets), Serbia (23, 015 tweets), Australia (12, 761 tweets), Nigeria (12, 038 tweets), South Africa (11, 715 tweets), Colombia (8, 270 tweets). As shown in Figure. 9, we generated a bar chart to present the more exact value based on the COVID-related tweets number of selected sample countries.

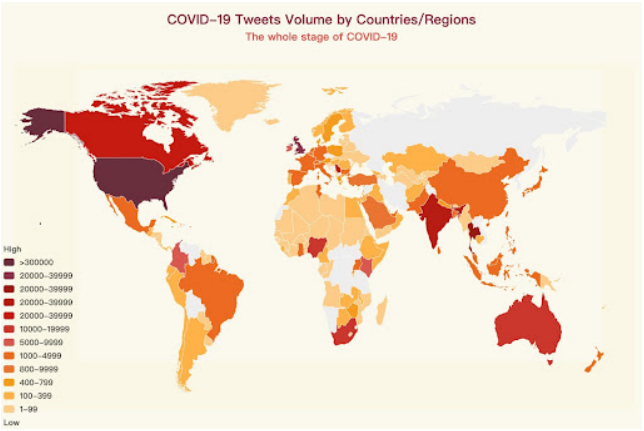


Figure 8. Tweets count of each country

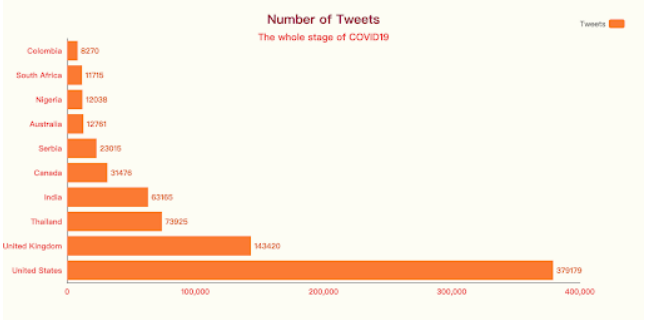


Figure 9. Number of tweets of each country

6.2 Sentiment analysis

We used the VADER to do sentiment analysis for selected ten countries with the most tweets and calculated the average sentiment scores for each country in the three stages. The results are shown in Table. 6.

Based on the result, we formed three heat maps. The heat maps as shown in Figure. 10(a) ˜ 10(c) reveals that the early stage's average sentiment scores are relatively high, with the highest score of 0.1809961(the United Kingdom). In contrast, the middle stage's average sentiment scores are relatively lower, with the lowest score of -0.0395131(the United Kingdom). And the late stage's average sentiment scores are in the middle, with the highest score of 0.1623043(Canada) and the lowest score of 0.0146162(Thailand).

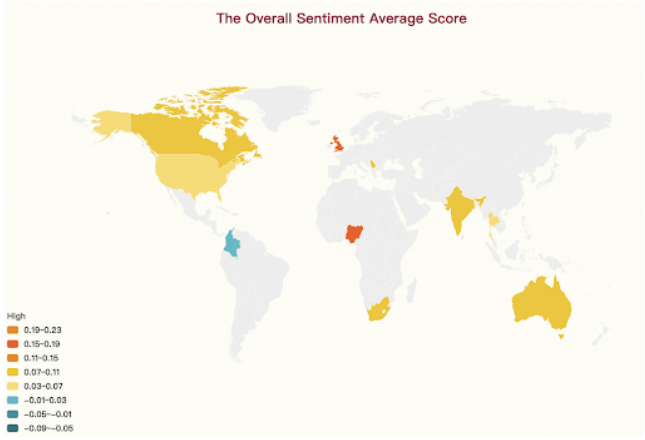


Figure 10(a). Heat map of the average sentiment score by country at the Early stage

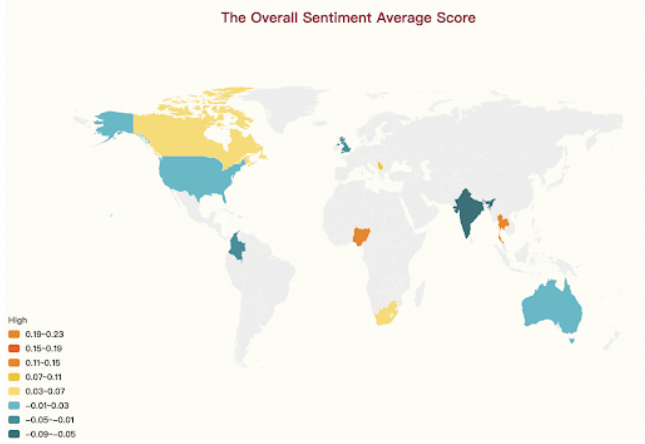


Figure 10(b). Heat map of the average sentiment score by country at the Middle stage

After obtaining the sentiment scores, we labeled tweets positive, negative, or neutral and counted the proportion of positive, negative, and neutral tweets across the three periods to know people's attitudes in different regions. We generated bar charts for each country, and the results are shown in Figure. 11(a) ˜ 11(c). We can see that positive tweets

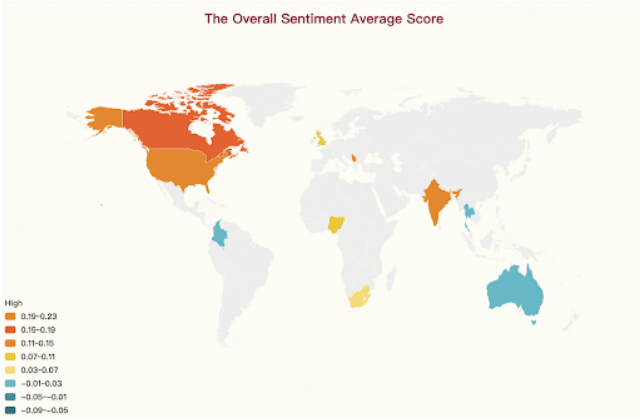


Figure 10(c). Heat map of the average sentiment score by country at the Late stage

account for a large proportion. But the differences are minor in most countries; for example, the proportions of positive, negative, and neutral tweets in Canada are 35.05%, 35.17%, and 29.78% at the early stage, 37.15%, 34.39%, and 28.46% at the middle stage, and 38.92%, 34.58% and 26.50% at the late stage. However, the differences are pronounced in some countries; for example, the related proportions in Nigeria are 49.41%, 24.86%, and 25.73% at the early stage, 44.75%, 29.71%, and 25.54% at the middle stage, and 43.09%, 28.46% and 28.46% at the late stage.

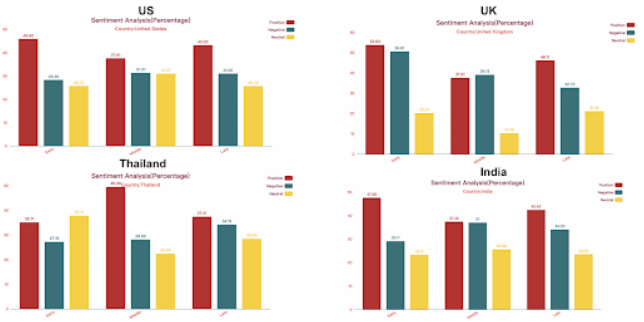


Figure 11(a). Sentiment analysis by countries

In order to know the overall sentiment scores and the sentiment trend among ten countries, we also calculated the average score, the maximum score, the minimum score, and the standard deviation for each period as shown in Table. 7 and the proportions of all tweets in selected ten countries as shown in Figure. 12. The sentiment analysis results in the early and late stages tend to be more positive but the results in the middle stage tend to be more negative.

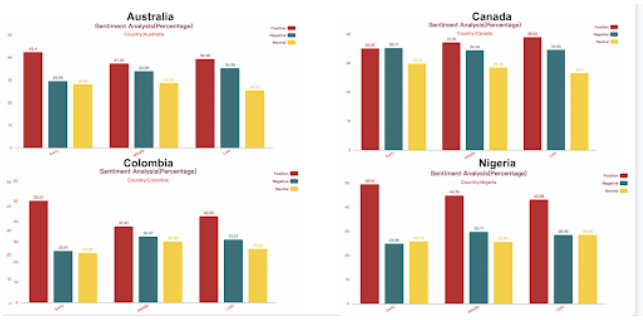


Figure 11(b). Sentiment analysis by countries

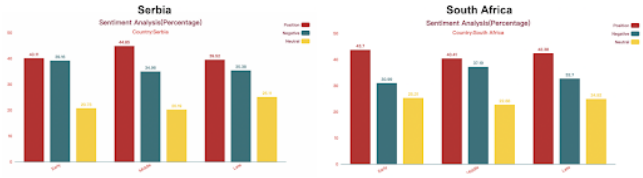


Figure 11(c). Sentiment analysis by countries

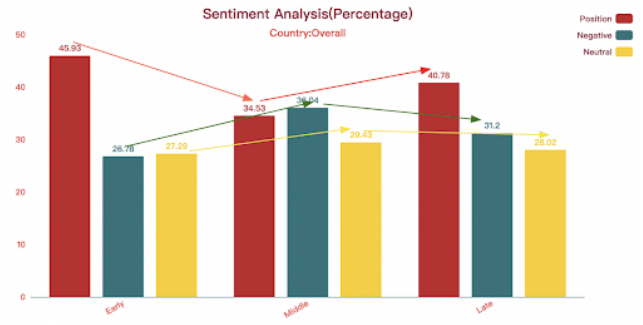


Figure 12. Overall sentiment analysis of all countries

6.3 Topic abstraction

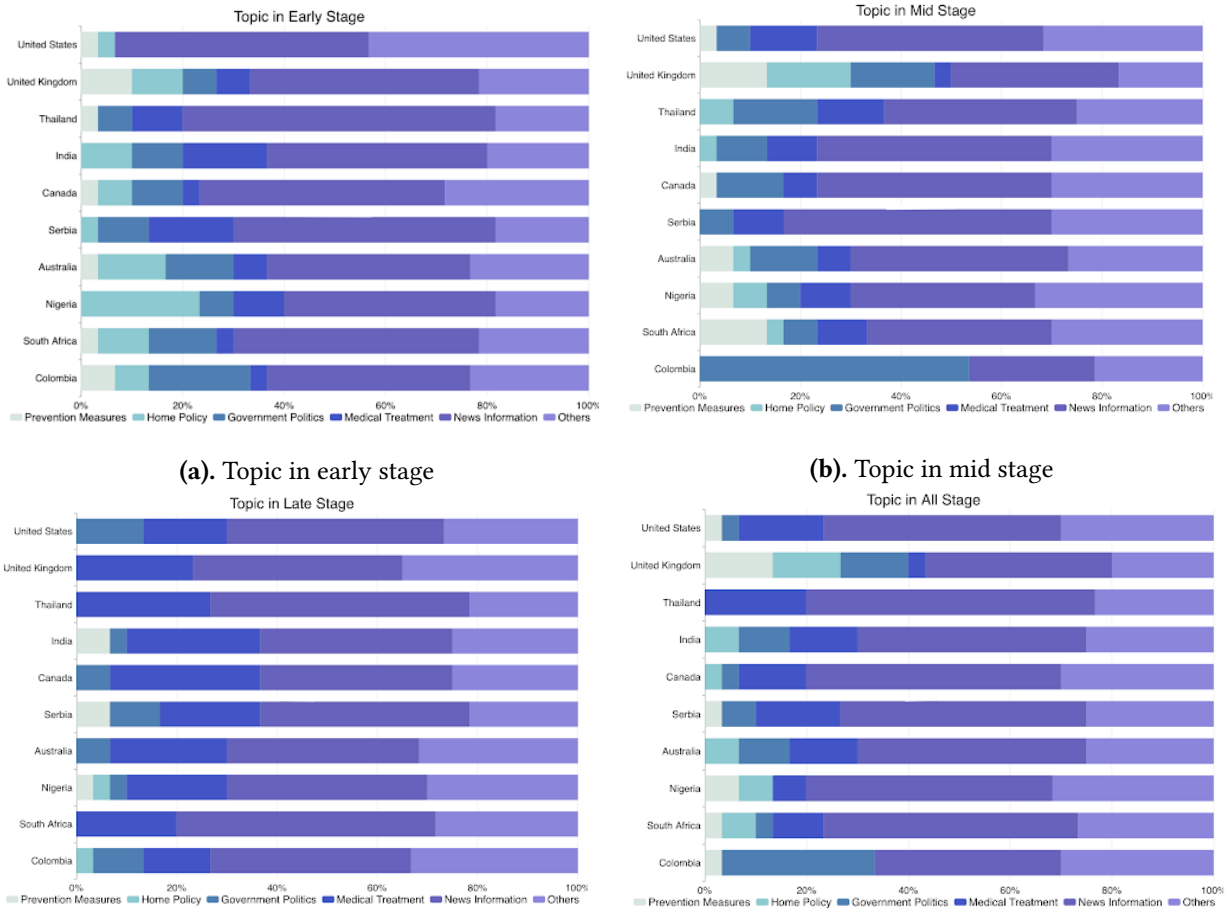
We splitted all the tweets into six categories by the LDA model, and visualized the final results for four stages, as shown in Figure. 13.

In addition, we also collected top-K salient terms during all the tweets, as shown in Figure. 14. Owing to the space limitation, the figure only shows top-10 salient terms with their percentage, so the probabilities of each term would not add up to 1.

7 Critical Discussion of Findings

7.1 Geographic data summary (Discussion engagement distribution)

Based on our statistical results of COVID-related tweets (Figure. 9), the United States is the most engaged country in debating the epidemic on Twitter. The reasons behind this include the USA has the most Twitter users (Figure. 15). Also, the massive amount of COVID confirmed cases and death



(a). Topic in early stage

(b). Topic in mid stage

(c). Topic in late stage

(d). Topic in all stages

Figure 13. The visualization results of four stages

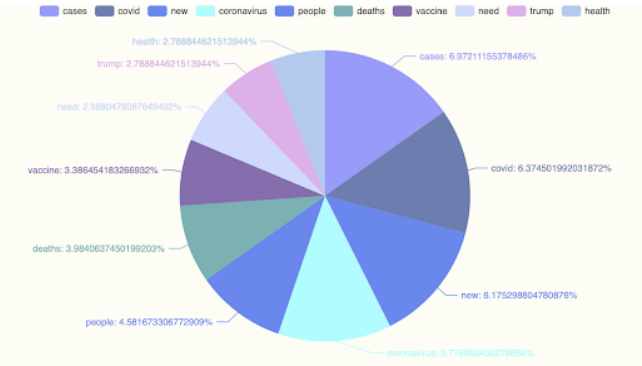


Figure 14. Top-10 Salient Terms

cases (Figure. 16) might also be attributed to the American partition.

In addition, we found some interesting observations of other countries by comparing the number of Twitter users (Figure. 15), published by the company Statista and our statistical results of COVID-related tweets (Figure. 9).

1. The United Kingdom has the third most COVID-related tweets volume through the development of COVID, whereas its number of Twitter users ranks in the top six.
2. Though Thailand is not the country with the most COVID-related tweets, its engagement degree is also relatively high. Although Thailand is only 10th in the number of Twitter users, its number of tweets about the COVID ranks third.
3. Same as Thailand, Canada is even below the top ten in terms of the number of Twitter users, but it's in the top five in terms of the number of COVID-related tweets

7.2 Sentiment analysis

Based on the heat maps as shown in Figure. 10(a) ˜ 10(c), the public's sentiments are more positive in the ten countries during the early and late stages but are more negative during

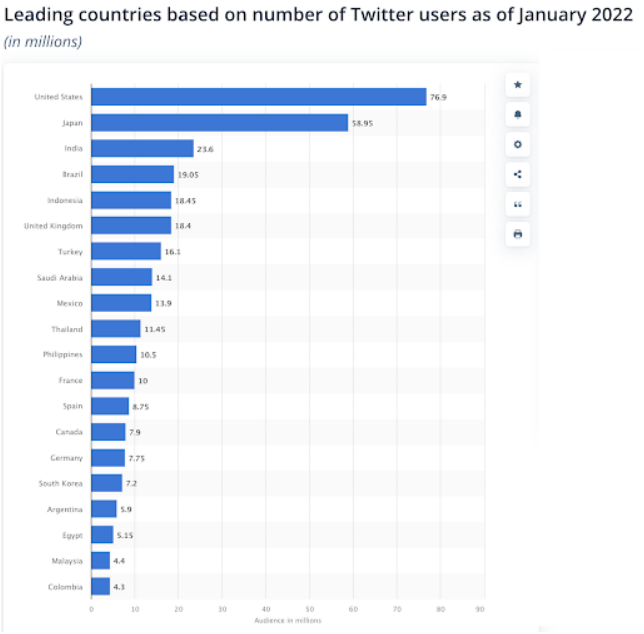


Figure 15. Number of Twitter users in USA

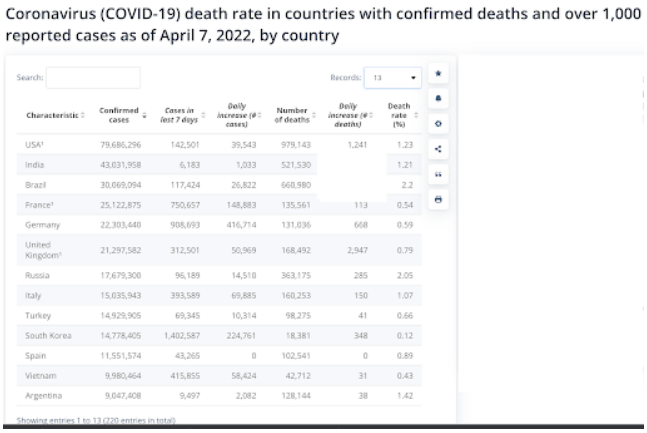


Figure 16. Number of COVID cases of each country

the middle stage. And from the Figure. 12, the overall sentiment trend also tends to decline in positivity from the early to the middle stage but tends to increase from the middle to the late stage. The reasons behind this result are that the epidemic has not seriously affected society at the early stage when the public showed more positive attitudes. But as time went on, in the middle stage, some variant viruses emerged and caused increasing infection numbers and death cases which has led to governments having started to take more strict prevention measures such as the lockdown policy to avoid the wider spread of the COVID19. Also, the vaccination number was small in the meantime. So, people's lives have been seriously affected during the middle stage. The public's sentiment tended to be a bit negative. However, in the late stage of the epidemic, although there were still many cases globally, people have adopted the inconvenience caused by this health crisis, such as the quarantine or wearing masks. And the vaccination rate was increasing hugely, which has been achieved 60.4%worldwide [22]. So, many countries have started to coexist with the epidemic to enable people can live as they did before the epidemic. As a result, there was a rebound in positivity at the late stage.

Besides, as shown in Figure. 11(a) ˜ 11(c), we found that in countries like the United States, the United Kingdom, India, etc., positive tweets decrease from the early stage to the middle stage but increase from the middle stage to the late stage, which is consistent with the overall sentiment trend (Figure. 12). In contrast, in some countries like Thailand and Serbia, the number of positive tweets increases first but decreases later. The factors that contribute to this situation may be that the epidemic outbreak happening in countries was different. Take the United States and Thailand as examples, Figure. 17 shows the number of COVID19 cases [23], and Figure. 18 shows the sentiment trend in the USA and Thailand. In the USA, the number of COVID19 cases began to increase rapidly in Oct 2020 (in the early stage), But in Thailand, the number of COVID19 cases started to proliferate in June 2021 (in the middle stage), which indicates that the epidemic began to spread earlier in the United States. So, the trends of the public's attitudes are different.

7.3 Topic abstraction

As shown in Figure. 13, the largest portion is the news information category in the whole stage, which may be caused by a large number of media accounts in our data set. The minuscule portion is the prevention measures category. Two reasons may be associated with this consequence. On the one hand, people may not believe in the role of equipment, such as the mask. This phenomenon is evident in Nigeria, where they hardly discussed the topic of masks in the early stage. On the other hand, people may not want to talk about masks, because they did not realize the severity of the COVID-19 in the early stage.

In contrast, in the mid-stage, the topic of prevention measures in some countries has increased significantly, proving the previous guess. People began to pay attention to preventive measures.

In the late stage, there is a considerable increase in the number of medical treatment topics. The damage caused by the COVID-19 may contribute to this phenomenon. However, the topic of stay-at-home policy has almost disappeared in some countries. Combined with the facts, this may be due to the herd immunization policies of some countries, such as the United Kingdom, as well as the relaxation of the outgoing policy.

8 Future Work

8.1 Limitation

Data Inconsistency.
The time distribution of the final dataset we used for analysis is uneven over the period because the initial dataset we obtained related to mid-epidemic covers only covers four days of COVID-related tweets, which causes a significant limitation on the time diversity of the dataset and affects the statistics of some countries' tweets number. Our search question RQ2 aimed to analyze the dataset based on the timeline of COVID development. So, the dataset we expected to obtain should have a time horizon of two years, from the end of 2019 to the beginning of 2022. We attempted to retrieve the data using Twitter API to collect a suitable dataset. However, Twitter has limited the response volume rate (180 tweets for version 1 Twitter API and 450 tweets for version 2 Twitter API) to requests per 15-minute window. Using this method to collect the dataset is time-consuming and not feasible for us due to the limited semester period. To expand the time diversity of our dataset, we searched related datasets from Kaggle and found several useful datasets. Though the currently available dataset does not fully achieve our expectation of time coverage and the limitation of our project still exists, the final dataset contains tweets from the epidemic's early, middle, and late stages. We can gain some insights regarding our research question.

Data Incompleteness.
We only select English tweets in our project for the analysis and filter out amounts of non-English tweets so that there is a lack of data in some countries such as Russia, etc.

Data Inaccuracy.
The source address of some English tweets may not be representative of that country. For example, some tourists are traveling in Japan. However, the native language of Japan is not English, so these English tweets sent in Japan do not accurately represent the opinion of the Japanese

8.2 Further improvement

The previous studies of COVID-19 tweets may suffer from dataset limitations. In the future, we will probably collect reliable data directly from the developer API provided by Twitter. To ensure the accuracy of the analysis, we may further interview several residents of each country.

For the sentiment analysis, we may further adopt the ensemble learning method to combine these two models considering that VADER and TextBlob specialize in different areas. Also, VADER cannot distinguish irony and cannot identify jargon or abbreviation phrases used in daily life, such as LOL, which means laugh out loud. We may also combine some machine learning models such as SVM to improve the accuracy.

For the topic abstraction, the LDA model cannot automatically generate a "Topic Name" for each cluster and further distinguish sub-topic. We may find and learn alternative algorithms with high performance and scalability.

Figure 17. The COVID 19 case trend: The United States vs. Thailand

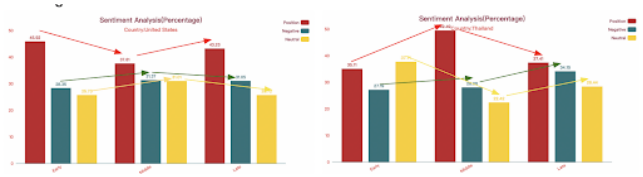


Figure 18. The Sentiment trend: The United States vs. Thailand

9 Distinguishment from similar projects

Our study focused on national level analysis regarding the Covid-19 among 219 countries, involved sentiment trend analysis throughout the development of the pandemic, statistics about the degree of participation distribution, and discovered public concerns. Apart from our study, some groups share the same research topic with us. Group 11 aimed to analyze the research questions based on the time dimension, and Group 6 aimed to analyze the tweets topics related to the "Work From Home."

10 Conclusion

This paper investigated how the discussion engagement regarding the Covid-19 pandemic was distributed globally in 219 countries, the sentiment changed throughout the development of the Covid-19 pandemic in different regions, and discovered the public concerns in the Covid-19 pandemic in the different regions.

We reported the tweet number distribution in 219 countries and hypothesized the reason behind the results. We showed the sentiment trend in the top ten countries with the most tweets. The analysis of sentiment scores shows that the public's sentiments are more positive in the ten countries during the pre and post-pandemic. At mid-epidemic, there is a decrease in positive percentage. We found that the trend of sentiment changes in line with the influence caused by

the development of the Covid-19 pandemic, for example, the vaccination and death rate.

We showed the main topics abstracted from the tweets in different regions, which is valuable and helpful for the government and society to discover public concerns in a global pandemic.

11 Contribution Statements

ZHANG, Cao (20803018)

In this study, I participated in the discussion and determination of research questions and I contributed one literature review of related work that inspired us on topic abstraction using LDA model.

For data collection, I contributed to the searching for useful datasets for post-pandemic. For data preprocessing, I contributed to the clean text and generate final CSV files. Also, I integrate the filtered dataset

For research questions, I contributed to the finding discussion for all research questions and visualization of the topic abstraction by LDA model. Finally, I contributed to the report structure and writing for the proposal, progress, and final report.

LU, Zijun (20784286)

In this study, I participated in the discussion and determination of research questions and I contributed one literature review of related work that inspired us on geographic analysis on a national level for discussion engagement on Twitter.

For data collection, I contributed to the searching for useful datasets by trying to use Twitter API to acquire stream tweet data. For data preprocessing, I contributed to the filter-non English tweets.

For research questions, I contributed to the finding discussion for all research questions and visualization of the tweets number distribution. Finally, I contributed to the report structure and writing for the proposal, progress, and final report.

HUANG, Xi (20785694)

In this study, I participated in the discussion and determination of research questions and I contributed one literature review of related work that inspired us on sentiment trend analysis and the method of using the VADER model.

For data collection, I contributed to the searching for useful datasets on Kaggle for pre-pandemic. For data preprocessing, I contributed to the mapping of location information to the uniform country name.

For research questions, I contributed to the finding discussion for all research questions and visualization of sentiment analysis results. Finally, I contributed to the report structure and writing for the proposal, progress, and final report.

12 Code and Dataset Availability

Codes for data cleaning and processing are available at: https://github.com/ZH0N9/CSIT-6000K-Course-Project-Social-Media-and-Epidemic-Modeling-Corona-

The datasets used in this research are available at: https://hkustconnect-my.sharepoint.com/:f:/r/personal/zlubc_connect_ust_hk/Documents/SocialNetwork_Dataset?csf=1&web=1&e=np0aaR

References

[1] Bruns, a., and weller, k. (2016). "twitter as a first draft of the present: and the challenges of preserving it for the future," in proceedings of the 8th acm conference on web science (hannover), 183–189. doi: 10.1145/2908131.2908174.

[2] Leonardo tortolero-blanco, d., 2020. social media influence in the covid-19 pandemic. scielo brazil. available at: https://www.scielo.br/j/ibju/a/nV6DpnQf7GWYrd94ZcHQBWz/?lang=en.

[3] Neha puri, eric a. coomes, hourmazd haghbayan, keith gunaratne social media and vaccine hesitancy: new updates for the era of covid-19 and globalized infectious diseases hum vaccine immunother (2020), pp. 1-8.

[4] Liu y, whitfield c, zhang t, hauser a, reynolds t, anwar m. monitoring covid-19 pandemic through the lens of social media using natural language processing and machine learning. health inf sci syst. 2021 jun 25;9(1):25. doi: 10.1007/s13755-021-00158-4. pmid: 34188896; pmcid: Pmc8226148.

[5] Han x, wang j, zhang m, wang x. using social media to mine and analyze public opinion related to covid-19 in china. international journal of environmental research and public health. 2020; 17(8):2788. https://doi.org/10.3390/ijerph17082788.

[6] Liu s, liu j. public attitudes toward covid-19 vaccines on english-language twitter: A sentiment analysis. vaccine. 2021 sep 15;39(39):5499-5505. doi: 10.1016/j.vaccine.2021.08.058. epub 2021 aug 17. pmid: 34452774; pmcid: Pmc8439574.

[7] Han x, wang j, zhang m, wang x. using social media to mine and analyze public opinion related to covid-19 in china. international journal of environmental research and public health. 2020; 17(8):2788. https://doi.org/10.3390/ijerph17082788.

[8] Liu s, liu j. public attitudes toward covid-19 vaccines on english-language twitter: A sentiment analysis. vaccine. 2021 sep 15;39(39):5499-5505. doi: 10.1016/j.vaccine.2021.08.058. epub 2021 aug 17. pmid: 34452774; pmcid: Pmc8439574.

[9] Cdc museum covid-19 timeline. david j. sencer cdc museum. (2022). retrieved 16 april 2022, from https://www.cdc.gov/museum/timeline/covid19.html.

[10] Timeline of the covid-19 pandemic. (n.d.). wikipedia. retrieved april 17, 2022, from https://en.wikipedia.org/wiki/Timeline_of_the_COVID-19_pandemic.

[11] Coronavirus (covid19) tweets. (n.d.). kaggle. retrieved april 17, 2022, from https://www.kaggle.com/datasets/smid80/coronavirus-covid19-tweets.

[12] Covid19 tweets. kaggle.com. (2022). retrieved 16 april 2022, from https://www.kaggle.com/gpreda/covid19-tweets.

[13] Sna_data. accessed 1 april 2022 from https://hkustconnect-my.sharepoint.com/personal/euhaq_connect_ust_hk/_layouts/15/onedrive.aspx?ga=1&id=%2Fpersonal%2Feuhaq%5Fconnect%5Fust%5Fhk%2FDocuments%2FSNA%5Fdata.

[14] Omicron - covid19 variant tweets. kaggle.com. (2022). retrieved 16 april 2022, from https://www.kaggle.com/datasets/shivamb/omicron-covid19-variant-tweets.

[15] 1.9m+ covid-19 tweets. kaggle.com. (2022). retrieved 16 april 2022, from https://www.kaggle.com/datasets/oktayozturk010/19m-covid19-

tweets.

[16] Omicron rising. kaggle.com. (2022). retrieved 16 april 2022, from https://www.kaggle.com/datasets/gpreda/omicron-rising.

[17] Marco lui and timothy baldwin. 2012. langid.py: An off-the-shelf language identification tool. in proceedings of the acl 2012 system demonstrations, pages 25–30, jeju island, korea. association for computational linguistics.

[18] C. hutto and e. gilbert, "vader: A parsimonious rule-based model for sentiment analysis of social media text", icwsm, vol. 8, no. 1, pp. 216-225, may 2014.

[19] Mohammad rezwanul huq, ahmad ali and anika rahman, "sentiment analysis on twitter data using knn and svm" international journal of advanced computer science and applications(ijacsa), 8(6), 2017. http://dx.doi.org/10.14569/IJACSA.2017.080603.

[20] Blei d m, ng a y, jordan m i. latent dirichlet allocation[j]. the journal of machine learning research, 2003, 3: 993-1022.

[21] Alteryx community "getting to the point with topic modeling | part 3 - interpreting the visualization.", 18 oct. 2021, https://community.alteryx.com/t5/Data-Science/Getting-to-the-Point-with-Topic-Modeling-Part-3-Interpreting-the/ba-p/614992.

[22] H. ritchie et al., "coronavirus pandemic (covid-19)", our world in data, 2022. [online]. available: https://ourworldindata.org/covid-vaccinations?country=OWID_WRL [accessed: 24- may- 2022].

[23] "github - cssegisanddata/covid-19: Novel coronavirus (covid-19) cases, provided by jhu csse", github, 2022. [online]. available: https://github.com/CSSEGISandData/COVID-19.[Accessed:18-May-2022].

A Appendix

Category	Example Words
Prevention Measures	PPE (Personal Protective Equipment), FaceMask, Wear, Distancing, etc.,
Home Policy	Lockdown, Stay, Home, Community, WorkAtHome, etc.,
Government Politics	Govt, UKchange, Trump, USA, Election, etc.,
Medical Treatment	NHS (National Health Service), Symptom, Vaccine, Hospital, etc.,
News Information	Cases, Unemployment, Pandemics, Breakingnews, Deaths, etc.,
Others	Military, Frightening, Post, Staff, Christmas, etc.,

Table 1. Example keywords of each category

		Predicted Condition	
		Positive	Negative
Actual Condition	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Table 2. Definition of Confusion Matrix

Measure Metrics	Formula
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Precision	$\frac{TP}{TP+FP}$
Recall	$\frac{TP}{TO+FN}$
F1-Score	$2 * \frac{Precision*Recall}{Precision+Recall}$

Table 3. Example keywords of each category

	Total Accuracy Rate
VADER	68.57%
Textblob	62.86%

Table 4.1. Total Accuracy Rate: VADER vs. Textblob

	Precision	Recall	F1-Score
Positive	0.64	0.86	0.74
Neutral	0.83	0.61	0.71
Negative	0.38	0.45	0.42
Accuracy			0.69
Macro Avg.	0.62	0.64	0.62
Weighted Avg.	0.72	0.69	0.69

Table 4.2. Classification Report of VADER

	Precision	Recall	F1-Score
Positive	0.51	0.78	0.62
Neutral	0.79	0.54	0.65
Negative	0.67	0.55	0.60
Accuracy			0.63
Macro Avg.	0.66	0.62	0.62
Weighted Avg.	0.68	0.63	0.63

Table 4.3. Classification Report of Textblob

Accuracy	Precision	Recall	F1-Score
0.78	0.77	0.76	0.76

Table 5. Evaluation Result of LDA

	Average Score (Early)	Average Score (Middle)	Average Score (Late)	Average Score (Total)
The United States	0.0671479	-0.0004464	0.1100290	0.0589102
The United Kingdom	0.1809961	-0.0395131	0.0841807	0.0752212
Thailand	0.0410057	0.1280469	0.0146162	0.0612229
India	0.1029816	-0.0876938	0.1450417	0.0534431
Canada	0.0730001	0.0387546	0.1623043	0.0913530
Serbia	0.0921577	0.0770727	0.1334588	0.1008963
Australia	0.0807203	0.0143533	0.0222502	0.0391080
Nigeria	0.1578143	0.1189309	0.0833877	0.1200443
South Africa	0.1071124	0.0496668	0.0539912	0.0702568
Colombia	0.0132534	-0.0371171	0.0229683	-0.0002985

Table 6. Average sentiment scores by countries

	Early Stage	Middle Stage	Late Stage
Average	0.0916190	0.0262055	0.0832228
Maximum	0.1809961	0.1280469	0.1623043
Minimum	0.0132534	-0.0876940	0.0146162
Standard deviation	0.0500074	0.0702516	0.0539838

Table 7. Sentiment scores among countries in the three stages