Progress Report of Social Media and Epidemic Modeling (Corona)

Team 3

Team Member

ZHANG, Cao	20803018
LU, Zijun	20784286
HUANG, Xi	20785694

1. Report Outline

This report presents the progress of our project, including the achievements of the current stage and the plan for the following phases.

At this stage, we have accomplished 35% of our project. We mainly focus on data collection, integration, and pre-processing. In the beginning, we make a COVID development timeline. As currently, it is post-pandemic; we divide the COVID development period into three stages: early, middle, and late. We collect the dataset based on different stages from the timeline and structure the dataset according to the time and country. After preprocessing the data, we count the number of tweets for each country. From the statistics results, we analyze the total number of tweets across multiple countries, contributing to the completion of RQ1 ***(Which region is the most engaged in debating the epidemic on Twitter?)***. We also explore and investigate the two models (***VADER and LDA),*** which will be applied to our project in the next phase.

The plan for the following phrases will be more focused on the two models. We will use two models to investigate RQ2 ***(How do people's attitudes to the epidemic change over time in the different regions?)*** and RQ3 ***(What topics are most concerned about when discussing the epidemic in different regions?)***, and further detail will be discussed in Section 5.

2. Introduction,

The severity of the COVID-19 has plunged people into widespread anxiety. And the public has become more and more sensitive toward the news and significant social events related to the epidemic. Due to the rise of social media in this era, Twitter has become one of the most extensively used social media worldwide, so it is common to know users' viewpoints by analyzing their tweets. Social media has both advantages and disadvantages in helping society overcome the pandemic [1]. One of the disadvantages of its use is the cause of infodemic [2].

Information diffusion is fast via social media, same as misinformation and rumor. 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.

Moreover, the government needs to stabilize public sentiments on influential social media during a hard time. 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 [4]. 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 [5]. And they also may know which country's policy is worth learning from by comparing public sentiments of different regions.

Also, 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.

As the motivation discussed above, we proposed three research questions:

- 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?

We believe those research questions reflect how people deal with the epidemic, which will help society develop better measures and predict future disease outbreaks.

3. Project progress and finding

3.1 COVID-19 Timeline

Here is the timeline including some significant events [6] during the COVID-19, as *Figure.1* shows.

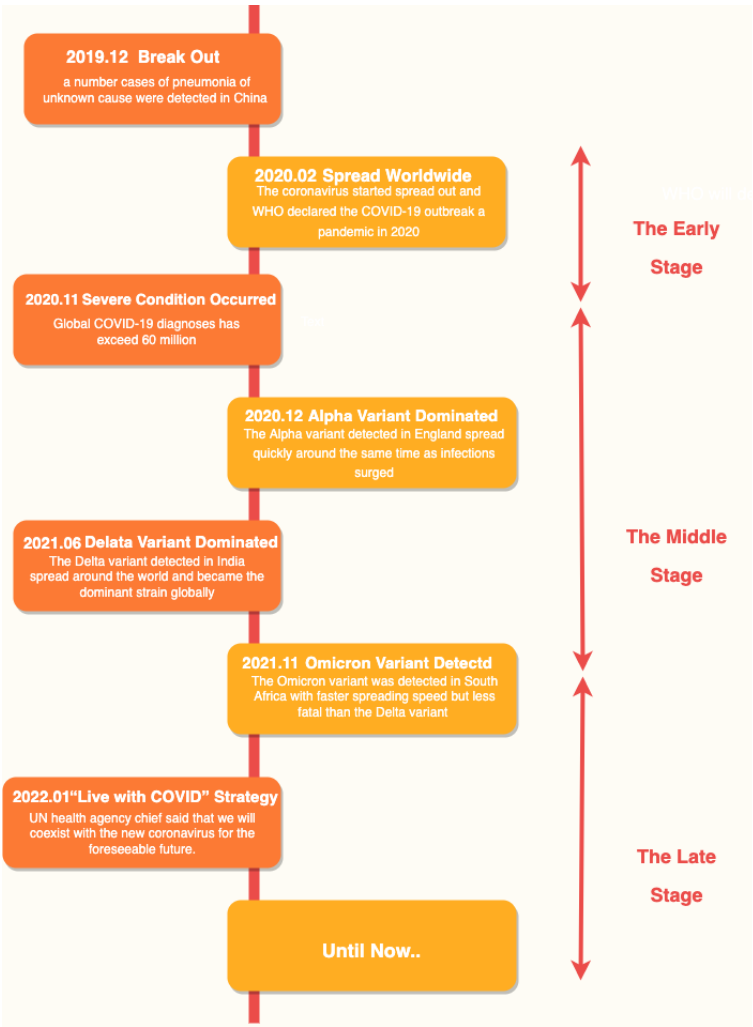


Figure. 1: Timeline of the COVID-19

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 infections cases and 6.19 million deaths globally [7].

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

(1). The early stage is from January,2020, when the virus started spreading worldwide, to November,2020.

(2). The middle stage is from December,2020, when the Alpha variant became dominant, to October 2021.

(3) The late stage is from November 2021, when the Omicron variant was detected, until now.

3.2 Data Collection

As mentioned in Section 3.1, we divided the data sets into three periods: early, middle and late stage. The early-stage dataset consists of two different datasets, one from March to April 2020 [8] and another one from July to August 2020 [9]. Both datasets are sourced from Kaggle, a website providing numerous open-source datasets. The middle-stage datasets will only contain the dataset containing data from November to December 2020 [10] provided by this course. The late-stage dataset is derived from 3 datasets, one containing data from December 2021 only [11], one containing the data from December 2021 to January 2022 [12], and the last one containing data from February to March 2022 [13]. However, there are still some limitations of our datasets, and we will discuss that part in Section 6.

However, since our datasets are collected from different places, the attributes of each dataset may not be similar, such that one of them has the column "user_url", while the others do not. To minimize the missing data of columns, we only keep the common columns to 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 problem is that although most datasets may have the column "verified users", they are not uniformly named. We also keep the naming uniform. Further, we leverage 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 will contain 13 columns such as "country_code", "user_id" and etc. We mainly use five of these columns: "user_id", "user_name", "country_code", "tweet_created_at", and "tweet" in this project.

Finally, we generate three folders to represent the three different stages. Each folder will contain CSV files for several different countries and a JSON file to store the number of tweets for each country. At the same time, we also write the number of tweets of each country before the file name of the corresponding country's CSV file. For example, "179064_United States.csv" indicates 179,064 tweets in the United States in that period.

3.3 Data Preprocessing

There may be unreliable and incorrect data in datasets, so we need to preprocess the data to avoid affecting our analysis.

Since Twitter users come from different countries, tweets may contain multiple languages. We mainly analyze English tweets in this project, so we filter out duplicate tweets and non-English tweets by langid [14], a standalone language identification tool. We further removed irrelevant information, such as URL links.

Since the research focus of the two models used in this project is different, the further preprocessing results would also be different. In RQ2, 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 will keep the full tweet. In RQ3, the LDA model is more focused on the potential concern of users, so information such as punctuation, emoji, numbers, retweet tag and stop words will be filtered out to preserve the central theme better.

3.4 Finding – Tweets number distribution

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. 2* 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). We generated a bar chart to present the more exact value based on the COVID-related tweets number of selected sample countries, as shown in *Figure. 3*.

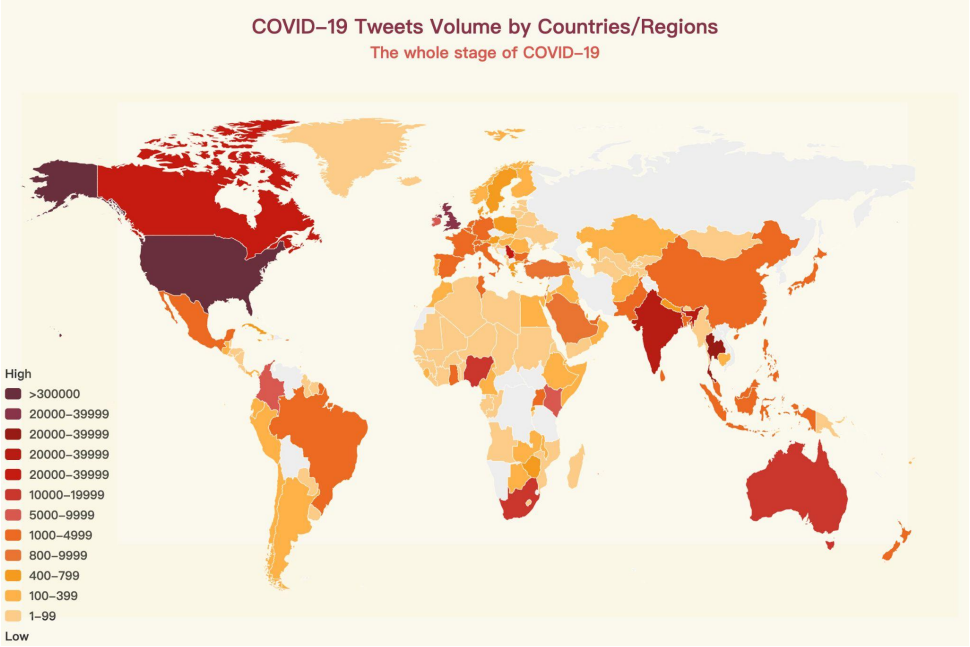


Figure. 2: Tweets count of each country

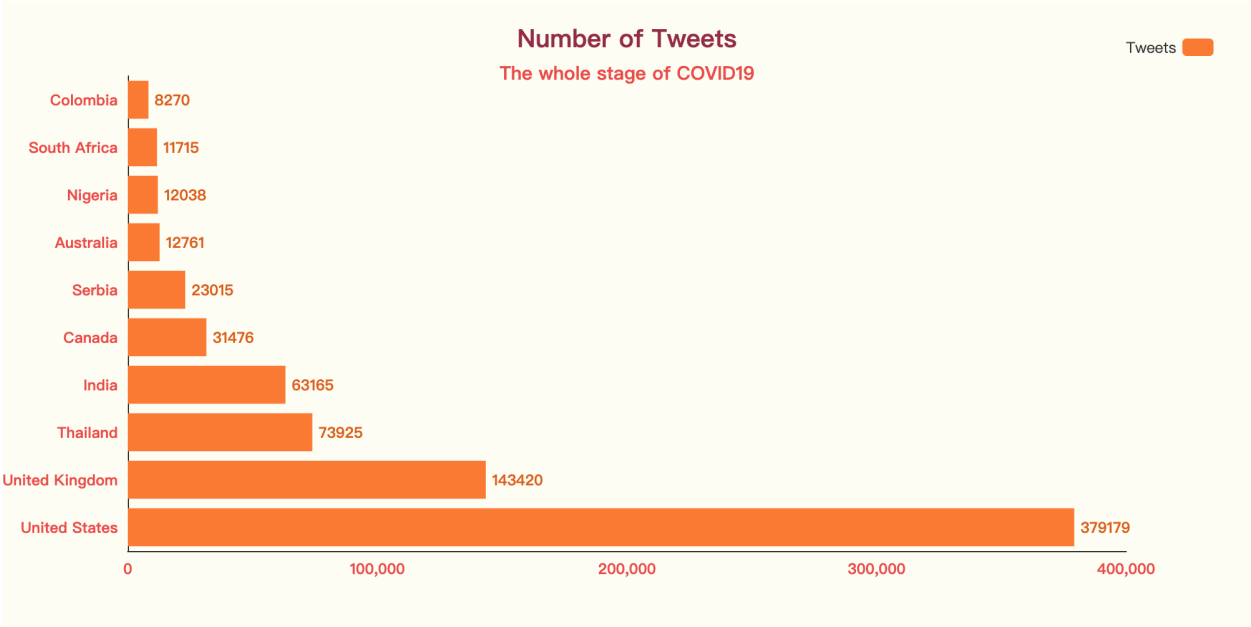


Figure. 3: Number of tweets of each country

Based on our statistical results of COVID-related tweets *(Figure 3)*, 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. 4)*. Also, the massive amount of COVID confirmed cases and death cases *(Figure. 5)* might also be attributed to the American partition.

In addition, we find some interesting observations of other countries by comparing the number of Twitter users *(Figure 4)*, published by the company Statista and our statistical results of COVID-related tweets *(Figure 3)*.

- Though Japan has the second-highest number of Twitter users, it seems like the COVID-related topic is not too popular in Japan. The number of COVID-related tweets in Japan is below the top 10.
- 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 6.
- 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 3rd.
- Same as Thailand, Canada is even below the top 10 in terms of the number of Twitter users, but it's in the top 5 in terms of the number of COVID-related tweets

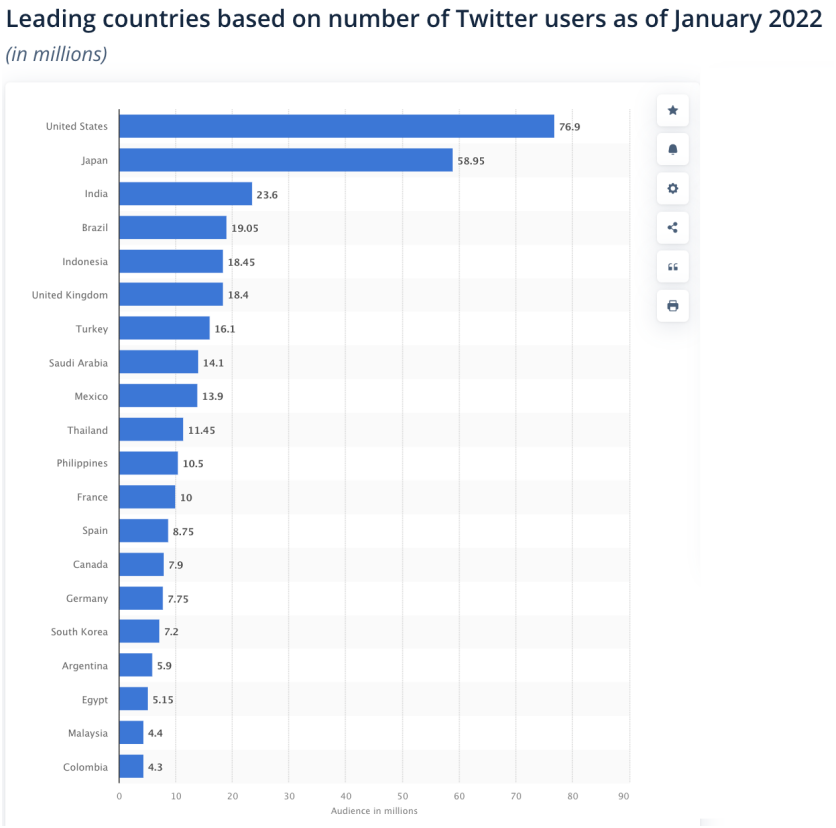
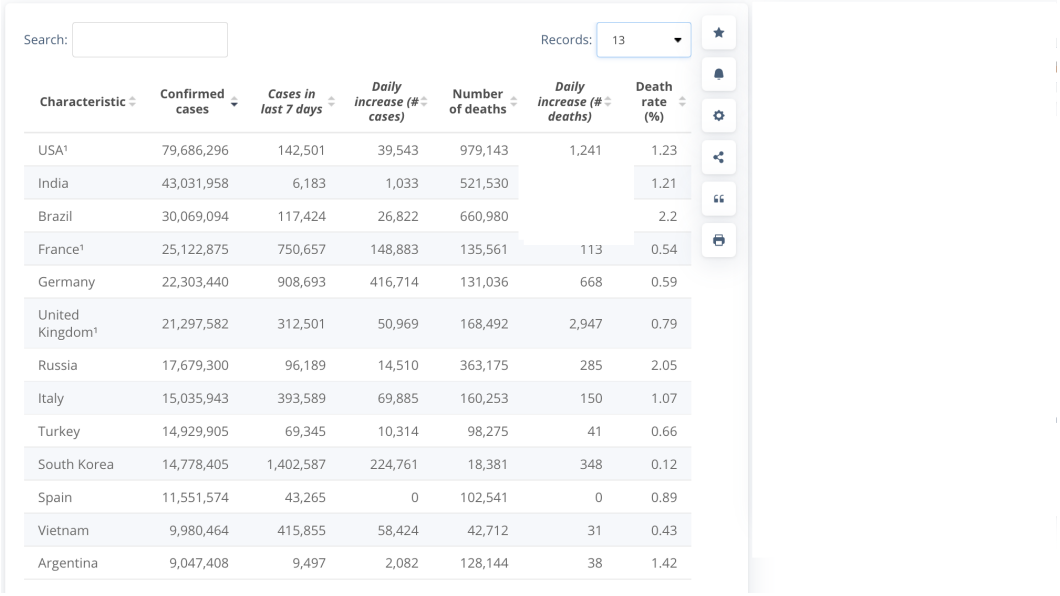


Figure. 4: [Number of Twitter users in USA]

Coronavirus (COVID-19) death rate in countries with confirmed deaths and over 1,000 reported cases as of April 7, 2022, by country



Search: Records: 13

Characteristic	Confirmed cases	Cases in last 7 days	Daily increase (# cases)	Number of deaths	Daily increase (# deaths)	Death rate (%)
USA[1]	79,686,296	142,501	39,543	979,143	1,241	1.23
India	43,031,958	6,183	1,033	521,530		1.21
Brazil	30,069,094	117,424	26,822	660,980		2.2
France[1]	25,122,875	750,657	148,883	135,561	113	0.54
Germany	22,303,440	908,693	416,714	131,036	668	0.59
United Kingdom[1]	21,297,582	312,501	50,969	168,492	2,947	0.79
Russia	17,679,300	96,189	14,510	363,175	285	2.05
Italy	15,035,943	393,589	69,885	160,253	150	1.07
Turkey	14,929,905	69,345	10,314	98,275	41	0.66
South Korea	14,778,405	1,402,587	224,761	18,381	348	0.12
Spain	11,551,574	43,265	0	102,541	0	0.89
Vietnam	9,980,464	415,855	58,424	42,712	31	0.43
Argentina	9,047,408	9,497	2,082	128,144	38	1.42

Figure. 5: Number of COVID cases of each country

4. The problem encountered

As our research questions state, we aim to explore how the people's attitudes and discussion engagement change towards the covid on Twitter across different regions. The location analysis is essential to our research. Information provided by geotagging on each tweet enables us to extract the value of longitude and latitude. However, according to the previous studies [15], the location service defaults off to Twitter users. Approximately only 0.85 % of tweets are geotagged, which means after filtering tweets, only a relatively small amount of the dataset will be available for us to analyze.

Due to this reason, instead of using geo-related information from tweets to map the users' location, we are using location information from users' profiles to gain location information. Although the consequences caused by this change include the reduction of the accuracy and representativeness of the dataset due to factors that multiple regions or cities share the same name, a certain level of accuracy and reliability is ensured by using user-level location information. The accuracy rate is predicted to be 76 % at the country level [16].

5. Plan for the next phase

5.1 Further plan

For the RQ2 ***(How do people's attitudes to the epidemic change over time in different regions)***, we will apply the VADER model, which is confirmed to be practical for our project, to analyze tweets for the top 10 countries with the most tweets in each stage. After that, we will generate the sentiment score and visualize the sentiment result to see the percentage of positive, negative, and neutral tweets in each country. And we can get to know the trend of the public attitude toward COVID-19 as time goes by in different regions.

For the RQ3 ***(What topics are most concerned about when discussing the epidemic in different regions?)***, we will use the LDA model to analyze the user's real concerns. Similar to research question 2, only the top 10 countries in each stage would be evaluated. We will generate 5 topics for each country and 10 words for each topic. The results would be visualized for easy observation.

5.2 VADER for attitude analysis

The Valence Aware Dictionary for sEntiment Reasoner(VADER) is based on the lexicon method [17]. Compared with other machine-learning-based methods such as SVM [18], there is no need to train the model and use the previously seen texts to determine the sentiments of new texts in a lexicon method. The working process of the VADER aims to map words to sentiments in the pre-built dictionary, where the intensity of emotion is expressed as the sentiment score. The sentiment score of a text or sentence is a compound of sentiment scores of all words(positive when compound value>=0.05, neutral when -0.05 <compound value<0.05, and negative when compound value<=-0.05), as *Figure. 6* shows.

Figure. 6: VADER workflow

Although the result with the machine learning method can be more accurate under a greater volume of training data, we intend to use the VADER to analyze the sentiments expressed in tweets. The reason is that first, the number of tweets data for each country is limited, so it may be hard to get the expected result by training model. Second, the length of the tweet text without a context is short, so it would be better for us to use a rule-based model after balancing the

efficiency and the accuracy. And third, tweet texts may contain lots of emojis, which has become a usual way for people to express emotions on social networks. The VADER can also process emojis which can improve the accuracy of sentiment results.

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) from the related works [19].

We try a sentiment test based on one dataset, including 30 tweets, to check whether it is practical. And the result is presented as the *Figure. 7* shows:

(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%.

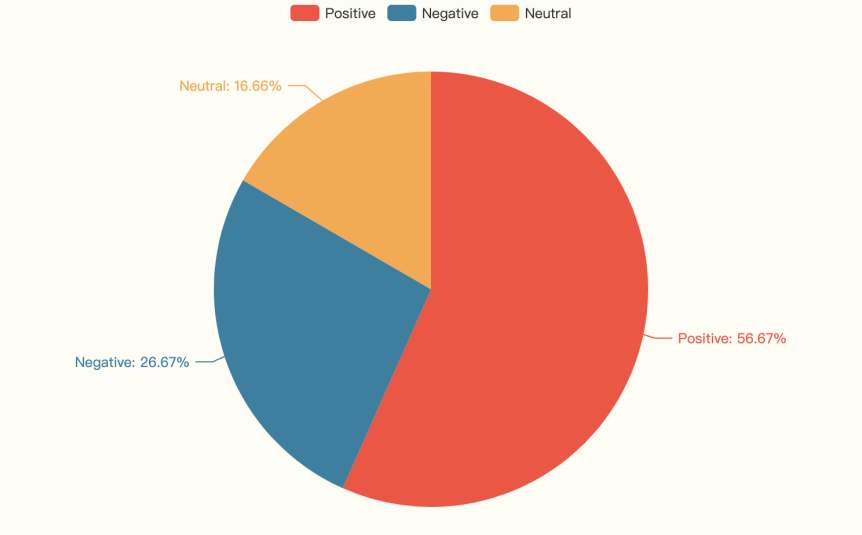


Figure. 7: Test Sentiment Result

We do some sentiment analysis by ourselves and compare the results. We find that around 73.33% of the result is the same as expected. And we will apply it to our analysis-needed data and see whether the performance is good or not.

5.3 LDA for topic extraction

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 argues 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.

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 is considered as our topic extraction tool.

In the LDA model, the target document is represented by a bag of words. 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 an a priori hyper-parameter 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 given by the user.

We choose 32 tweets from Armenia in the early stage to test the LDA model and define the number of topics as 5. *Figure.8* shows topic 1, with its most relevant words.

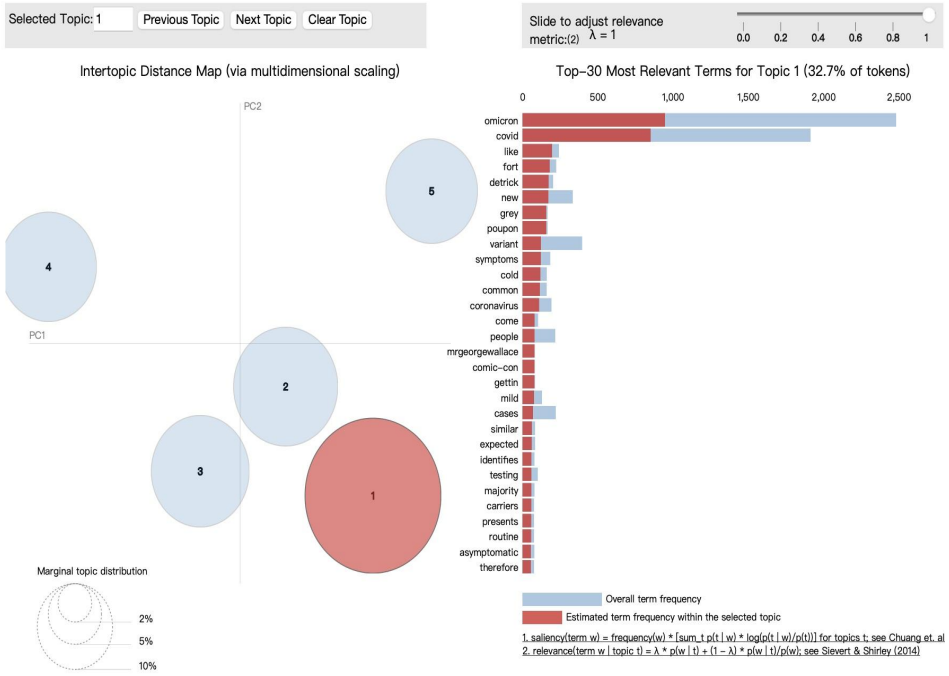


Figure. 8: Visualization of tweets from Armenia

6. Limitation

6.1 Data Inconsistency

The time distribution of the final dataset we are using for analysis is uneven over the period, because the initial dataset we obtained related to mid-epidemic covers only 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 aims to analyze the dataset based on the timeline of COVID development. So, the dataset we expected to obtain has 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.

6.2 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.

7. Summarization

In conclusion, we did data collection, integration, and pre-processing at the current stage. We further extract the top 10 countries with some statistical analysis. Owing to the world's largest population of Twitter users and the highest number of COVID confirmed cases, we find that the USA has become the most engaged country in discussing COVID-related topics.

In the final report, we plan to use the VADER model to analyze the sentiments on the tweets and find out the trend of the public attitude. Likewise, the LDA model is another strategy for us to extract the most popular topics of each country.

8. Reference

[1] 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>

[2]Cinelli, M., Quattrociocchi, W., Galeazzi, A. *et al.* The COVID-19 social media infodemic. *Sci Rep* 10, 16598 (2020). https://doi.org/10.1038/s41598-020-73510-5

[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]: 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

[5]: 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.

[6]: CDC Museum COVID-19 Timeline. David J. Sencer CDC Museum. (2022). Retrieved 16 April 2022, from https://www.cdc.gov/museum/timeline/covid19.html.

[7]: 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.

[8]: Coronavirus (covid19) Tweets. (n.d.). Kaggle. Retrieved April 17, 2022, from https://www.kaggle.com/datasets/smid80/coronavirus-covid19-tweets.

[9]: COVID19 Tweets. Kaggle.com. (2022). Retrieved 16 April 2022, from https://www.kaggle.com/gpreda/covid19-tweets.

[10]: 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.

[11]: Omicron - Covid19 Variant Tweets. Kaggle.com. (2022). Retrieved 16 April 2022, from https://www.kaggle.com/datasets/shivamb/omicron-covid19-variant-tweets.

[12]: 1.9M+ COVID-19 Tweets. Kaggle.com. (2022). Retrieved 16 April 2022, from https://www.kaggle.com/datasets/oktayozturk010/19m-covid19-tweets.

[13]: Omicron Rising. Kaggle.com. (2022). Retrieved 16 April 2022, from https://www.kaggle.com/datasets/gpreda/omicron-rising.

[14] 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.

[15]Sloan, L., & Morgan, J. (2015). Who Tweets with Their Location? Understanding the Relationship between Demographic Characteristics and the Use of Geoservices and Geotagging on Twitter. *PloS one*, *10*(11), e0142209. https://doi.org/10.1371/journal.pone.0142209.

[16]Mahmud, J., Nichols, J. and Drews, C., 2014. Home Location Identification of Twitter Users. *ACM Transactions on Intelligent Systems and Technology*, 5(3), pp.1-21.

[17]: Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

[18]: 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.

[19]: Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. Proceedings of the International AAAI Conference on Web and Social Media, 8(1), 216-225. Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/14550.

[20]: Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. the Journal of machine Learning research, 2003, 3: 993-1022.