CSE 6242 Project Final Report

Shengyun Peng, Junyan Mao, Guanchen Meng,
Zefang Liu, Yuxuan Wang, Huili Huang
{speng65,jmao,gmeng9,liuzefang,ywang3336,hhuang413}@gatech.edu

1 INTRODUCTION

Coronavirus disease 2019 (COVID-19) is a contagious disease that has taken numerous lives. It is important for health professionals to monitor the trend and distribution of such a pandemic in order to distribute medical resources accordingly. Social media chatters may serve as a good data source for this task because there is a strong correlation between COVID-19-related talks and actual COVID cases. Twitter, as a leading social platform, generates over 800 million tweets in a single day, and recently a majority of discussions are focusing on COVID-19, which makes Twitter an ideal data source for our project. Current COVID-19-related websites only provide visualizations on daily cases. The relationship between COVID-19 and Twitter chatters has not been fully explored yet. Therefore, in this project, our goal is to evaluate the relationship between the COVID-19 cases and the number/content of tweets in each country, and provide data visualization to assist users in making decisions and designing guidelines for work and school during the pandemic. To be specific, we would like to demonstrate how COVID-19 spread across the world and manifest the correlation between trend in tweets and the actual trend in COVID-19. Besides, a word cloud of most frequently occurring tokenized phrases is presented to give users a brief overview of the most popular topics regarding COVID-19 on Twitter.

2 LITERATURE SURVEY

Trajkova et al. [18] presented a study of Twitter retweet frequencies during the period the COVID-19 epidemic. Their findings can be considered to facilitate our understanding of Twitter information spread. However, the study does not discuss methods for altering Twitter information flows. Al-Rakhami et al. [1] conducted an analysis on information credibility of Twitter posts during the COVID-19 pandemic. These models can be used for COVID-19 information refinement in our project. This study can be improved by considering complex tweets with news and emotional contents. Rufai et al. [12] demonstrated the increasing of Twitter usages by the majority of G7 world leaders on the subject of COVID-19. Such information from the G7 leaders can be analyzed in our project. But, the safety and accuracy of the tweets should be paid attention. Samant et al. [13] developed a flexible interactive dashboard for COVID-19 data visualization with their demographics based on multi-layer network. The event-based and parameter-based aggregate analysis and visualization can be used for comparisons of features and consequences in the COVID-19 data and as an improvement for their study. Amin et al. [2] proposed an intelligent model for COVID-19 pandemic detection from Twitter messages based multiple machine models. This dataset can be incorporated into our project. However, more categories can provide fine-grained datasets. Torres et al. [17] analyzed the URL sources being posted by influential Twitter accounts, which can help us comprehend the information dissemination of the COVID-19 tweets. Their Twitter data can be extended to contain epidemiology and vaccinations questions.

Xue J. et al. [20] proposed a machine learning algorithm that uses a Twitter dataset to understand discussions related to the pandemic. It provides us some ideas of techniques when analyzing our own dataset. However, we plan to generate an interactive interface to present our analyzing results. Clement et al. [6] built an interactive dashboard to display real-time global trends of COVID-19, from which our team could learn the packages and tools for data visualization. But our team would elaborate on that by interpreting our dataset to extract more information before displaying it to the users. Kwan et al. [14] discussed a framework constructed to analyze general public's emotions and sentiments in response to COVID-19 based on their tweets. Its topic has some overlaps with our project as we both use tweets as the data source. However, we would build an interactive interface to visualize our analysis results. Signorini et al. [15] proposed that Twitter could be a reliable data source for both tracking disease level and analyzing public's reactions to a pandemic. Although it

talked about H1N1 instead of the COVID-19, our team could learn what hidden information we might obtain from our dataset. Rufai et al. [12] analyzed word frequency pattern and conducted sentiment analysis on tweets posted by the public and WHO. Although not decided yet, our group may include word frequency pattern as part of our analysis. But besides that, we would also focus on visualizing the data and the result of analysis. Thomas et al. [16] built a neural network to predict users' location based on the Twitter data. It might not be helpful to our project as our dataset should already include the geological information. Also, besides location, we will include more features like time and frequency.

Burton et al. [5] evaluated the location information provided by Twitter's Global Positioning System for infoveillance and infodemiology to prevent overrepresenting or underrepresenting. As Karami et al. [8] stated, the number is less than 1% now while the number of users has increased by 90 million required an update. Nekliudov et al. [11] studied the impact of COVID-19 pandemic on mental health. This paper expounded a potential function of our project, monitoring COVID-related social anxiety. One drawback is the generalizability of this paper, since they only sampled from Russian. Li et al. [10] conducted a survey on social media use in China during the COVID-19 pandemic. This study is geographically complementary to our research. They also stated that the aggregated social media had more influence than public media in China as a potential improvement. Skunkan et al. [4] analyzed public recognition to COVID-19 pandemic by data mining. This research may provide validation to our analysis results. However, they only collected data from English language twitters not representing the general situation. Xu et al. [19] used the same database, the geotagged tweets to measure how many people had travelled since the pandemic of COVID-19. This article inspires us the correlation between data statistics and corresponding state laws. Kim et al. [9] studied the influence of media on people's information obtaining and decision making. They pointed out that keywords and filters may biased people's understanding of truth, such as facts about COVID-19. Our project may improve this phenomenon by explicitly showing the discussion popularity.

3 PROPOSED METHOD

3.1 Dataset

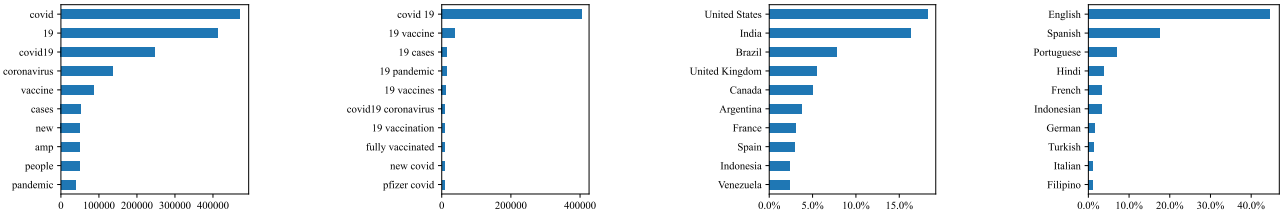
In this project, we use two datasets: COVID-19 daily cases [7] and COVID-19 Twitter chatter dataset [3]. COVID-19 Data is provided by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University, which has detailed COVID-19 cases each day, month and year. Twitter chatter dataset is a public dataset maintained by the Panacea Lab at Georgia State University. It consists of daily tweets acquired from the Twitter stream related to COVID-19 chatter and provides top 1000 frequent terms, bigrams, and trigrams for Natural Language Processing (NLP) tasks. It's important to mention that the original dataset only contained "tweet_id"s, which are unique identifiers used by Twitter to identify tweets. We utilize the "Tweepy" library, which serves as a wrapper around Twitter's API, to hydrate the raw data with specific geo-locations and its original text.

The frequent terms from the Tweet dataset are shown in Figure 1a with their counts. We can find the most frequent terms includes "covid", "19", "covid19", "coronavirus", and "vaccine". Most frequent terms are related to COVID-19 pandemic and vaccines, which present people's concern about this world-wide health crisis. A bigram is a sequence of two adjacent elements. The frequent bigrams from the Tweet dataset are shown in the Figure 1b. Similar to the comment terms, these bigrams are close to the COVID-19 pandemic and vaccines. For distributions of tweets over countries and languages, the countries with the largest numbers of the tweets are shown in the Figure 1c. The most tweets are from the United States, India, Brazil, United Kingdom, and Canada. These countries have large populations of Tweet users. Also, the languages with the largest numbers of the tweets are shown in the Figure 1d. The most tweets use English, Spanish, Portuguese, Hindi, and French.

3.2 Global COVID-19 Visualization

In this section, we mainly discuss steps to visualize the COVID-19 data in 3D format.

The first step is collecting COVID-19 data including daily cases, deaths and recovered. The CSSE raw data introduced in the previous section has the following schema: Province/State, Country/Region, Longitude,



(a) Counts of Terms **(b) Counts of Bigrams** **(c) Tweet Countries** **(d) Tweet Languages**

Figure 1: COVID-19 Related Tweets

Latitude and the daily COVID-19 cases. We can directly join the data with the global geo dataset, which has the 3D coordinates of each country in the world.

The second step is plotting all the countries on a sphere. Inspired by the Google WebGL Globe, we modify the code to display the Choropleth on the sphere. WebGL is a web component to represent data visualization layers on a 3D globe in a spherical projection. To obtain the coordinates of all the countries on the globe, we have adopted GeoJSON, which is a format for encoding a variety of geographic data structures. All the corresponding COVID-19 data will be displayed with a window when hovering on it.

Finally, we provide the dynamic visualization function. The user can easily visualize how the COVID-19 has grown starting from the pandemic breakout to the latest date. The animation is implemented by the slide bar and get value function. A slide bar displaying all the dates will be placed on the bottom. When the date proceeds, the get value function will use the current date on the slide bar as an input and fetch the corresponding COVID-19 case data on that day and attach it to the sphere with the correct color scheme. The larger the infected numbers, the darker the color selected.

The user interface also provides the exact numbers of total confirmed cases, deaths, and recovered. These numbers will change according to the date selected in the slide bar. Moreover, three corresponding line charts are also plotted on the webpage to track changes over long periods of time.

3.3 COVID-19 Tweet Visualization

In this section, we visualize the Twitter chatter dataset on a similar globe described in 3.2. We will be displaying the number of tweets, the most popular terms, along with some other useful information in a country over a time period.

To accomplish this, we first retrieve the full Twitter chatter dataset from [3], filter out records without geographic location, then partition the data into small blocks indexed by date and country. This arrangement provides major advantages when it comes to visualizing time-related data.

We select 6000 COVID-19 related tweets with geographic coordinates from 408312 tweets in the March and April 2021. The latitudes and longitudes of the tweet locations are used for tweet visualization as in the Figure 2. The height of one bar represents the number of tweets from one location, where the higher bar is for more tweets. Also, different colors are used for different numbers of tweets, where the red color represents more tweets. The total number of tweets is shown on the top left corner of the web page. We also listed the top-10 tweet countries and tweet languages with their counts of tweets as in the Figure 1c and 1d. The top-3 countries of the COVID-19 related tweets are United States, India, and Brazil. And top-3 languages are English, Spanish, and Portuguese.

3.4 COVID-19 Twitter Word Clouds

The relationship between COVID-19 related tweets focus and the severity of COVID-19 is studied. To make the results more visible, the Word Cloud are introduced to evaluate the frequently discussed topic. We implement d3-cloud plugin, which provides the layout, font size, rotation and other useful feature, to generate fancy word clouds. We generate word cloud during the COVID-19 second wave. From the world daily new confirmed COVID-19 cases collected by Our world in Data, we collect and combine the data date_top1000bigrams.csv from three time periods within the COVID-19 second

Figure 2: COVID-19 Tweet Globe

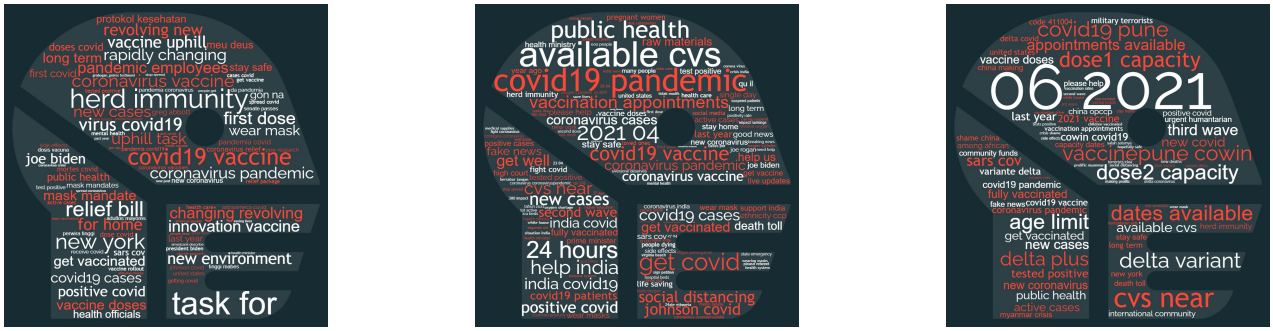
wave(from 2021-03 to 2021-07): [2021-03-02,2021-03-06], [2021-04-23,2021-04,27], and [2021-06-19,2021-06,23] illustrate the lowest daily case before the second wave, the highest daily cases in the second wave, and the lowest daily cases after the second wave, respectively. The topics are chosen based on the discussed frequencies in five-day periods centered at the peak and valleys of COVID-19 confirms in English-based contexts. The data is cleaned by filtering out non-English vocabularies by a python library named langdetect and deleting words with repetitive meanings such as "covid19" and "covid-19". All topics on the cloud were mentioned more than 1000 times. The font size are determined by the frequency of words appear in tweets. We choose red and white as the displayed word color to give a sense of urgency, and black as the background color to add solemnity. Three fonts are applied to words randomly to add visual variation to the graph. We also design the overall shape of the clouds to be a human head wearing a mask, which is a combination of the symbol of COVID-19 and an analogy of "what people are thinking."

The animation of monthly word clouds is generated to illustrate the variation of COVID-19 related tweets more interactively. We collect and combine the data date_top1000terms.csv for each month and filter out the top60 popular topics to create the word clouds. Same as the analysis of the COVID-19 topic during the second wave, all non-english words are filtered out. To combine with Tweet visualization in 3.3, we pick up the filtered dataset from Jan.2021 to Apr.2021 to generate the animation. The monthly dynamic word clouds are displayed with the Tweet Globe in 3.3 to explain the variation of the COVID-19 topic during this four months. We create a layout of frame named "Monthly Word Cloud from Jan. to Apr" on the left side of the Tweet Globe usring D3. The word cloud will be updated continuously from Jan.2021 to Apr.2021. In order not to bored the reviewers, different from the visualization of the second wave, colorful categorical colors category20 is implemented to enhance the diversity of the word clouds. To ensure the readability of the word cloud information, only 25-30 terms from the monthly dataset are picked up randomly each time to update the word cloud. In this way, reviewers are able to observe the changes of word cloud content overtime without see the repetitive information. Figure 6 shows the Twitter globe and the word cloud visualization.

4 EXPERIMENTS AND EVALUATIONS

To evaluate if our interactive COVID-19 global distribution map is user-friendly and informational, we examine it in the following aspects. The corresponding results are also displayed.

- Basic user interactions are implemented including rotating the globe, dragging the slide bar, clicking the button in the navigation bar leads to a new webpage. Besides, three line charts demonstrate the trend of the COVID-19 cases.
- Hovering on a country allows users to see country name, daily COVID-19 confirmed, deaths and recovered cases. This result is displayed in 4. We use France as an example. The selected date is Nov. 24th, 2021.
- Dragging along the slide bar displays daily COVID-19 case data and each country changes to corresponding color scheme. Meanwhile, the total cases are dynamically displaying on the right, as shown in 5. If the user clicks the play button, the animation will starts playing the evolution of COVID-19 with color changing among different countries.



(a) 2021 03-First Valley (b) 2021 04-Second Peak (c) 2021 06-Second Valley

Figure 3: Three simple graphs

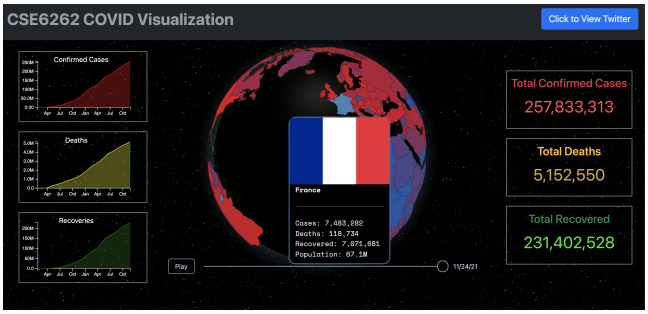


Figure 4: User interface of the global COVID-19 cases: 1) line charts, 2) total number, 3) country-wise cases of the selected date in the slide bar, and 4) the navigation bar

For the COVID-19 related tweets, we display bars with different heights for numbers of tweets from one geographic locations. We also list the number of tweets using one language and the number of tweets from one country. The user interface of the COVID-19 Twitter visualization is shown in the Figure 6. The following functions are inspected:

- Displaying bars on a earth globe to show the numbers of COVID-19 related tweets on their geographic locations, where a location with more tweets has a higher bar and a hotter color.
- Showing total number of COVID-19 related tweets in the 2021 March and April.
- Listing the most common countries for COVID-19 tweets in the dataset, where top 10 countries are presented with their numbers of tweets. The countries with the most tweets are listed on the top of the figure.

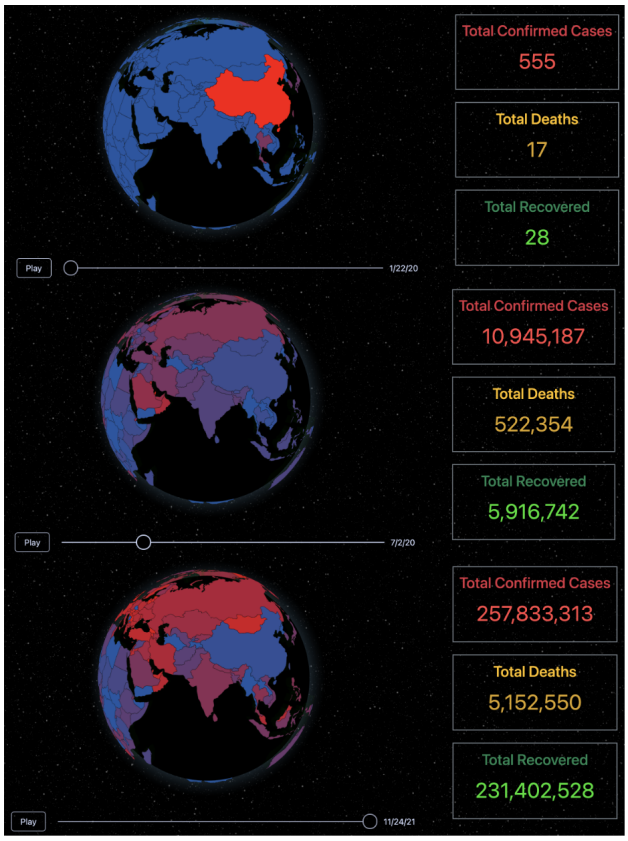


Figure 5: Dragging along the slide bar displays daily COVID-19 case data and each country changes to corresponding color scheme.

- Listing the most common used languages for tweets, where top 10 languages are shown with their numbers of tweets. The languages with the most tweets are listed on the top of the figure.

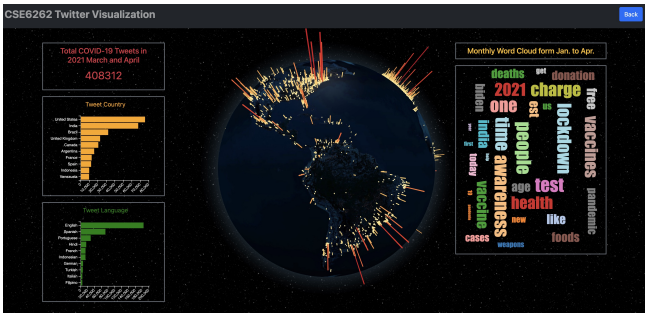


Figure 6: User interface of the COVID-19 Twitter visualization: 1) total number, 2) bar charts, 3) globe with tweet geographic locations, and 4) word cloud

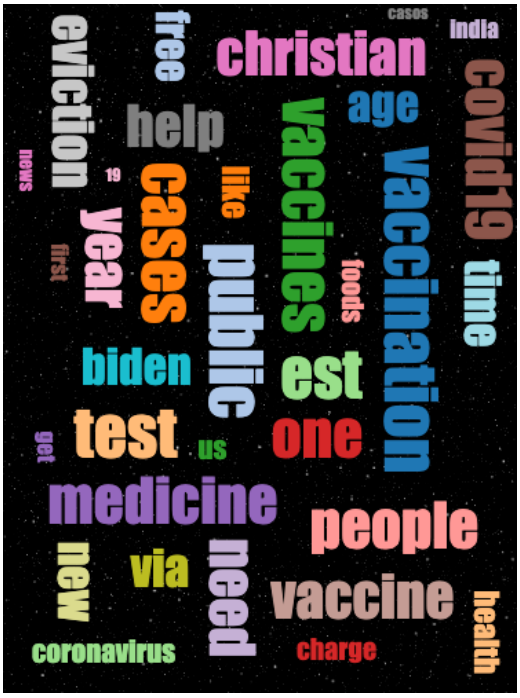


Figure 7: Animated word cloud of the monthly COVID-19 related topic

The daily word cloud visualization results reflect transitions towards the epidemic to some degree. From the word clouds during the second wave and 2021, we confirm that the public attention was positively correlated to the confirm numbers. As shown in Fig. 3, people become sensitive during the peak and use words like "people dying" and "life saving", while at valleys the topics are more likely to be general topic like "public health" and "Joe Biden". The way people fighting against COVID-19 is also changing. At the beginning, besides vaccines, people believed in herb immunity and wearing masks, but as the confirmed number rapidly grows, people dived into vaccination and talked about specific brands, doses, and appointments of vaccines. Besides, the location names appearing on the cloud often reflect the most heavily influenced districts in the world, such as “help Italy” and “help India”. Same trend is found for the word cloud analysis in 2021, when the epidemic becomes severe again (i.e., the third wave in July), bigrams like "getting worse" and "saving life" are shown as the popular topic. When the situation got better in August, bigrams like "get well" and "well soon" became the topic of discussion.

5 CONCLUSION

In this project, we implement a COVID-19 data and Twitter relationship visualizer. The 3-D global COVID-19 visualizer provides accurate and necessary information for all users. Users can easily know daily and total COVID-19 confirmed, deaths, and recovered cases from our website. Interactive functions like play button and slide bar also display the evolution of the pandemic. The geographic locations with more COVID-19 cases have more COVID-19 related tweets. The number of COVID-19 tweets from one location has a positive relation with its number of COVID-19 cases visually. The word cloud shown in Figure 7 indicates the potential popular COVID-19 terms in these location area, which further confirm the positive correlation between the COVID-19 tweets and the COVID-19 cases. The word cloud analysis of the COVID-19 related tweets during the second wave demonstrates the positive correlation between these tweets and the severity of COVID-19. For example, when the COVID-19 cases increased, the tweets terms such as the location information and medicine treatment always reflect some hints of the severity.

6 EFFORT DISTRIBUTION

All team members have contributed similar amount of effort. To be more specific, S.H. Peng and G.C. Meng finished the global visualization part. J.Y. Mao and Z.F. Liu finished Twitter geographic data visualization. H.L. Huang and Y.X. Wang finished Twitter word cloud construction.

REFERENCES

[1] Mabrook S. Al-Rakhami and Atif M. Al-Amri. 2020. Lies Kill, Facts Save: Detecting COVID-19 Misinformation in Twitter. *IEEE ACCESS* 8 (2020), 155961–155970. https://doi.org/{10.1109/ACCESS.2020.3019600}

[2] Samina Amin, M. Irfan Uddin, Heyam H. Al-Baity, M. Ali Zeb, and M. Abrar Khan. 2021. Machine Learning Approach for COVID-19 Detection on Twitter. *CMC-COMPUTERS MATERIALS & CONTINUA* 68, 2 (2021), 2231–2247. https://doi.org/{10.32604/cmc.2021.016896}

[3] Juan M. Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, Ekaterina Artemova, Elena Tutubalina, and Gerardo Chowell. 2021. A Large-Scale COVID-19 Twitter Chatter Dataset for Open Scientific Research—An International Collaboration. *Epidemiologia* 2, 3 (2021), 315–324. https://doi.org/10.3390/epidemiologia2030024

[4] Sakun Boon-Itt and Yukolpat Skunkan. 2020. Public Perception of the COVID-19 Pandemic on Twitter: Sentiment Analysis and Topic Modeling Study. *JMIR Public Health Surveill* 6, 4 (11 Nov 2020), e21978. https://doi.org/10.2196/21978

[5] Scott H Burton, Kesler W Tanner, Christophe G Giraud-Carrier, Joshua H West, and Michael D Barnes. 2012. "Right Time, Right Place" Health Communication on Twitter: Value and Accuracy of Location Information. *J Med Internet Res* 14, 6 (15 Nov 2012), e156. https://doi.org/10.2196/jmir.2121

[6] Frincy Clement, Asket Kaur, Maryam Sedghi, Deepa Krishnaswamy, and Kumaradevan Punithakumar. 2020. Interactive Data Driven Visualization for COVID-19 with Trends, Analytics and Forecasting. In *2020 24th International Conference Information Visualisation (IV)*. 593–598. https://doi.org/10.1109/IV51561.2020.00101

[7] Ensheng Dong, Hongru Du, and Lauren Gardner. 2020. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet infectious diseases* 20, 5 (2020), 533–534.

[8] Amir Karami, Rachana Redd Kadari, Lekha Panati, Siva Prasad Nooli, Harshini Bheemreddy, and Parisa Bozorgi. 2021. Analysis of Geotagging Behavior: Do Geotagged Users Represent the Twitter Population? *ISPRS International Journal of Geo-Information* 10, 6 (2021). https://doi.org/10.3390/ijgi10060373

[9] Yoonsang Kim, Jidong Huang, and Sherry Emery. 2016. Garbage in, Garbage Out: Data Collection, Quality Assessment and Reporting Standards for Social Media Data Use in Health Research, Infodemiology and Digital Disease Detection. *J Med Internet Res* 18, 2 (26 Feb 2016), e41. https://doi.org/10.2196/jmir.4738

[10] Xiaojing Li and Qinliang Liu. 2020. Social Media Use, eHealth Literacy, Disease Knowledge, and Preventive Behaviors in the COVID-19 Pandemic: Cross-Sectional Study on Chinese Netizens. *J Med Internet Res* 22, 10 (9 Oct 2020), e19684. https://doi.org/10.2196/19684

[11] Nikita A Nekliudov, Oleg Blyuss, Ka Yan Cheung, Loukia Petrou, Jon Genuneit, Nikita Sushentsev, Anna Levadnaya, Pasquale Comberiati, John O Warner, Gareth Tudor-Williams, Martin Teufel, Matthew Greenhawt, Audrey DunnGalvin, and Daniel Munblit. 2020. Excessive Media Consumption About COVID-19 is Associated With Increased State Anxiety: Outcomes of a Large Online Survey in Russia. *J Med Internet Res* 22, 9 (11 Sep 2020), e20955. https://doi.org/10.2196/20955

[12] Sohaib R. Rufai and Catey Bunce. 2020. World leaders' usage of Twitter in response to the COVID-19 pandemic: a content analysis. *JOURNAL OF PUBLIC HEALTH* 42, 3 (SEP 2020), 510–516. https://doi.org/{10.1093/pubmed/fdaa049}

[13] Kunal Samant, Endrit Memeti, Abhishek Santra, Enamul Karim, and Sharma Chakravarthy. 2021. CoWiz: Interactive Covid-19 Visualization Based On Multilayer Network Analysis. In *2021 IEEE 37TH INTERNATIONAL CONFERENCE ON DATA ENGINEERING (ICDE 2021) (IEEE International Conference on Data Engineering)*. IEEE, 2665–2668. https://doi.org/{10.1109/ICDE51399.2021.00299} 37th IEEE International Conference on Data Engineering (IEEE ICDE), ELECTR NETWORK, APR 19-22, 2021.

[14] Jolin Shaynn-Ly Kwan and Kwan Hui Lim. 2020. Understanding Public Sentiments, Opinions and Topics about COVID-19 using Twitter. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 623–626. https://doi.org/10.1109/ASONAM49781.2020.9381384

[15] Alessio Signorini, Alberto Maria Segre, and Philip M Polgreen. 2011. The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. *PloS one* (May 2011).

[16] Philippe Thomas and Leonhard Hennig. 2018. Twitter Geolocation Prediction Using Neural Networks. In *Language Technologies for the Challenges of the Digital Age*, Georg Rehm and Thierry Declerck (Eds.). Springer International Publishing, Cham, 248–255.

[17] Jorge Torres, Vaibhav Anu, and Aparna S. Varde. 2021. Understanding the Information Disseminated Using Twitter During the COVID-19 Pandemic. In *2021 IEEE INTERNATIONAL IOT, ELECTRONICS AND MECHATRONICS CONFERENCE (IEMTRONICS)*, Chakrabarti, S and Paul, R and Gill, B and Gangopadhyay, M and Poddar, S (Ed.). IEEE; Inst Engn & Management; IEEE Vancouver Sect; IEEE Toronto Sect; SMART; Univ Engn & Management, 418–423. https://doi.org/{10.1109/IEMTRONICS52119.2021.9422523} IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), ELECTR NETWORK, APR 21-24, 2021.

[18] Milka Trajkova, A'aeshah Alhakamy, Francesco Cafaro, Sanika Vedak, Rashmi Mallappa, and Sreekanth R. Kankara. 2020. Exploring Casual COVID-19 Data Visualizations on Twitter: Topics and Challenges. *INFORMATICS-BASEL* 7, 3 (SEP 2020). https://doi.org/{10.3390/informatics7030035}

[19] Paiheng Xu, Mark Dredze, and David A Broniatowski. 2020. The Twitter Social Mobility Index: Measuring Social Distancing Practices With Geolocated Tweets. *J Med Internet Res* 22, 12 (3 Dec 2020), e21499. https://doi.org/10.2196/21499

[20] Jia Xue, Junxiang Chen, Ran Hu, Chen Chen, Chengda Zheng, Yue Su, and Tingshao Zhu. 2020. Twitter Discussions and Emotions About the COVID-19 Pandemic: Machine Learning Approach. *Journal of Medical Internet Research* 22, 11 (2020). https://doi.org/10.2196/20550