CSE 6242 Project Proposal

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

1 HEILMEIER'S QUESTIONS & EXPECTED INNOVATION

1) ***What are you trying to do? Articulate your objectives using absolutely no jargon.***

COVID-19 has been a prevalent topic since 2019 on Twitter. As part of our normal daily social media, we get around 4.4 million tweets a day. We want to find out the relationship between the COVID-19 cases and number of tweets in each country.

2) ***How is it done today; what are the limits of current practice?***

There are several COVID-19 cases visualization websites nowadays. Most of them display the number of daily new cases and deaths like the official websites of CDC and New York Times. However, none of them contain the visualization of tweets related to COVID-19. [3] provides a dataset of tweets acquired from the Twitter Stream related to COVID-19 keywords (coronavirus, 2019nCoV). But no visualization is provided.

3) ***What's new in your approach? Why will it be successful?***

The relationship between COVID-19 and tweets has not been explored yet. It will be successful because we can obtain the COVID-19 cases dataset from CDC. And there is an existing huge dataset regarding the number of tweets, tweet contents from the pandemic outbreak. We will also incorporate machine learning algorithms like time series analysis to find out the trend and make predictions on what will happen in the future. Few visualization websites provide this kind of function, and this is the expected innovation of our project.

4) ***Who cares?***

EVERYONE cares! From the breakout of this pandemic, the working and living situations have been thoroughly changed. Everyone needs visualization data to make decisions and provide guidelines for work and school.

5) ***If you're successful, what difference and impact will it make, and how do you measure them?***

The project will demonstrate how COVID-19 spread across the world and manifest the correlation between trend in tweets and the actual trend in COVID-19. Besides, word cloud analysis provides the most popular topic regarding COVID-19 on Twitter. By comparing the our COVID-19 distribution map to New York Times [16], we can verify the accuracy of our visualization. We will send surveys to check if our website is user-friendly and provides useful information.

6) ***What are the risks and payoffs?***

Privacy concerns over getting users' geo-location data from Twitter are the main risks. The payoff for this project is enormous. Using the Twitter data, we can help health professionals predict upcoming spikes in cases so that they can distribute necessary medical resources and make policy adjustments accordingly.

7) ***How much will it cost?***

This project will cost $0 as we will use our local computing resources along with the free Amazon EC2 quota.

8) ***How long will it take?***

We are planning to finish the proposed project by Sunday, November 28, 2021.

9) ***What are the midterm and final "exams" to check for success? How will progress be measured?***

By "midterm", all the data are fetched, cleaned, and processed and the website will have some basic visualizations. By "final", all the visualizations and interactions proposed are added. We plan to hold weekly meetings to discuss progress and updates.

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.

[11] 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. [12] 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. [13] 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. [14] 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. [11] 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. [15] 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. [7] stated, the number is less than 1% now while the number of users has increased by 90 million required an update. Nekliudov et al. [10] 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. [9] 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. [8] 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 PLAN OF ACTIVITIES

All team members have contributed similar amount of effort. Time estimates: 1) Shengyun Peng and Junyan Mao finish data cleaning in 1 month, 2) Guanchen Meng and Zefang Liu finish basic visualization in 2 months and 3) Yuxuan Wang and Huili Huang finish webpage interactions and word cloud before submission.

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] 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

[8] 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

[9] 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

[10] 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

[11] 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}

[12] 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.

[13] 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

[14] 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).

[15] 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.

[16] New York Times. 2021. Coronavirus World Map: Tracking the Global Outbreak. Retrieved October 12, 2021 from https://www.nytimes.com/interactive/2021/world/covid-cases.html

[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