COVIZVAX: What Twitter Data Reveals About COVID-19 Vaccine Adoption

Fa Fa Ke, Abdullah Mobeen, Abhishek Sharma, Mushi Wang

Abstract—The coronavirus disease (COVID-19) has been raging on for more than 21 months now, claiming lives waves after waves. The pandemic has accounted for more than 4.3 million deaths globally [36], while the World Health Organization (WHO) has classified at least 4 variants of concern [22]. The only good news is that medical scientists have come up with multiple effective vaccines in record time. While vaccine supplies remain a challenge for most of the world, the United States of America (US) has secured enough doses to fully vaccinate their entire population. As COVID-19 cases have largely come under control in the country, thanks to the massive immunization efforts, we are seeing very high vaccine hesitancy in some states. We try to understand this issue by analyzing people's opinions using sentiment analysis and the things they discuss with topic modeling. To get an in-depth analysis on the Twitter datasets, we introduce COVIZVAX: an interactive tool that can assist in exploratory data analysis of the people's sentiments and important topics of tweets about COVID-19 vaccinations.

Index Terms—COVID-19, Vaccination, Visualization, Sentiment Analysis, Topic Modeling

1 Introduction

The US is a country where its residents have the freedom to choose whether or not they want to get vaccinated. Since each state in the US have different vaccination policies, and political parties preferences, people tend to have various opinions on vaccination. It is interesting if we could find out why an individual chooses to receive vaccination versus someone who decides not to. We take a stab at this question by analyzing the COVID-19 vaccine Twitter datasets and examine the public opinions. Our goal is to build a tool that allows users to monitor and explore public attitudes on COVID-19 vaccines. Through this tool, we help the users discover divergent attitudes on COVID-19 vaccines.

We will attempt to answer the following research questions:

1. How are the views on COVID-19 vaccines distributed across the US?

2. Can a visual tool be built to provide interactive overview of the views on vaccines?

Previous efforts to analyze social media data lie under two broad categories: (i) sentiment analysis and topic modeling on COVID-19 related tweets, and (ii) COVID-19 related tweets across geospatial and temporal distributions. For sentiment analysis, Muller et al. [21] developed and open-sourced a model named COVID-Twitter BERT, which outperforms some of the established baseline models for sentiment analysis on Twitter vaccine datasets. However, the scope of their research is limited to natural language tasks. To map COVID-19 related tweets to temporal and spatial distributions, DeVerna et al. [12] developed a dashboard that monitors tweets related to COVID-19 vaccines and maps them across time as well as the US's map. They, however, do not consider the sentiment of the tweets and rely on the popular hashtags.

Our approach is to extract geospatial and post content information from COVID-19 related Twitter posts. The next step is to filter the dataset to exclude entries with invalid location and texts. These filtered out posts includes non-existing cities, irrelevant information and any texts presented in a non-English language. Afterwards, we perform sentiment analysis and topic extraction to determine common themes and perspectives amongst the Twitter posts based on different locations and communities. Eventually, we visualize our findings in a graph and a map interactively.

Our project contains three major contributions. Firstly, we find out the similarities and differences of opinions on the topic of COVID-19 vaccinations relevant to the cities or states they reside in. Secondly, we try to extract topic keywords and identify trending topics geospecially and temporally. Lastly, we plan on building an interactive visualization to present our findings specifically enabling the viewers to zoom in and out and link tweets that have similar topics. In addition, more detailed information is displayed if a particular tweet is selected by the user. We use visualizations as a tool for performing exploratory data analysis to help uncover patterns and insights from tweets dataset.

2 Related Work

Social media has gained popularity in the last few years. People use Facebook, Twitter, Instagram, Reddit, and other social media platforms for consuming media and information. So naturally, other researchers have studied the effect social media has had when it comes to COVID-19.

2.1 COVID-19 Key Research Areas

One of the big areas of research has been the spread of information and misinformation on social media. In [28], Dean Schillinger et al. provide a framework to evaluate the positive and negative effects of social media information sharing. In [11], Matteo Cinelli et al. compared different social media platforms to understand how information spreads on these, but found no significant difference in the spread of information from reliable or unreliable sources. In [20], Alessandro Lovari discusses how the spread of misinformation online led to an erosion of trust in public institutions in Italy. It describes the Ministry of Health's efforts in mitigating the effects of misinformation, and suggests a coordinated multi-strategy involving various relevant institutes. In [15], Raquel G. Hernandez et al. find that less than 10% of the COVID-19 related tweets stemmed from the medical community. The authors call this "Health Care Provider Social Media Hesitancy".

Other studies have focused on how political leaders leveraged social media during COVID-19. In [26], Sohaib R. Rufai et al. look at G7 leaders' use of Twitter, and find that 83% viral tweets were "informative", 9% were "Morale-boosting" and 7% were "Political", with all "Political" tweets coming from Donald Trump. In [29], Kurt Sengul discussed how an Australian right-wing politician used Facebook for promoting her nativist policies.

Finally, there are studies that focus on understanding users' views on topics related to vaccinations, wearing masks, social distancing, and other COVID-19 related subjects. Hussain et al. analyzed public attitude towards vaccines by leveraging artificial intelligence [16]. Tan et al. looked at people's reaction towards physical distancing by diving into Facebook posts and comments [33]. Keller et. al examined how public health ministries can use social media comments to better promote wearing masks and prevent additional spread of COVID-19 [17]. By analyzing another perspective, Al-Ramahi et al. dived into the

topics that may influence people to not wear masks using machine learning techniques [3]. Last but not least, Ahmed et al. analyzed the social network of Twitter posts related to masks [2]. All of these papers discusses a general attitude and reactions on the topics of COVID-19 using different research methods.

2.2 Social Media Visualizations

Social media visualization research can be broadly divided into three major categories: (i) social networks (ii) spatial-temporal information, and (iii) text analysis. Social network visualizations are typically graphs consisting of nodes connected by edges where the nodes represent the users, and the edges are the relationships between the users derived from regular online interactions. Spatial-temporal data are encoded into visualizations using timestamps and location data collected from each user's operations such as posting and commenting on Twitter. This type of data can be presented with region or map associated visualizations. Lastly, text analysis involves analyzing the user generated contents within the Twitter posts. The analysis results can be communicated through word clouds, topic flow diagrams with interactive filters and search options [10].

2.2.1 Sentiment Analysis

There are a lot of text analysis modules in existence. Among the numerous social media platforms, Twitter is the most commonly used and popular platform for politicians and people to express political opinions. Moreover, Twitter has been the subject of much sentimental analysis research. Hence, in our study, we are going to focus on sentimental analysis on tweets that are related to COVID-19 vaccination. Sentiment analysis tasks include classification of sentiment polarity expressed in text (e.g., positive, negative, neutral), identifying sentiment target/topic, opinion holder identification, and identifying sentiment for various aspects of a topic, product, or organization [1]. Early visualization techniques use bar charts and heat maps to represent opinion mining and sentimental analysis results. More recent tools presents data in river flow, map, and tunnel-liked diagrams with colors classifying the different sentiments [18]. The approaches of sentiment analysis can be categorized into two types, supervised methods and lexicon-based methods.

Supervised methods are based often on Naive Bayes, Maximum Entropy and Support Vector Machines [27] with the combination of specific features, such as, word n-grams [7], Part-Of-Speech (POS) tags [6], and tweets syntax features(hashtags, retweets, etc.) with accuracies about $80-84\%$.

Lexicon-based methods use a pre-prepared sentiment lexicon which essentially maps each word to a sentiment score and aggregate the sentiment scores of all words in a certain document. For instance, Semantic Orientation CALculator (SO-CAL) [32], SentiWordNet [5], etc.

One obvious limitation is that sentiment scores are strictly based on the lexicon itself regardless of their contexts. On the other hand, sentiment scores are fully depend on the present of words or related to syntactical features.

2.2.2 Keywords Analysis

TweeVist and TweetViz both use word clouds to display keywords from Twitter posts analysis. The word cloud in TweeVist is generated based on the time span and classification of the user's geolocation based on the store location. [35] TweetViz focuses on displaying keyword and topic distribution generated from Twitter posts contents and hashtags. [30] Other than the interactive and dynamic visualizations, the analysis can also be static in the forms of creative infographics and matrix, bar graphs, heat maps, and line graphs displaying time series data with various categorization information. [14]

Another key area in keywords analysis is topic modeling. Methods of topic modeling can be categorized into 4 parts, (i) Latent Semantic Analysis (LSA), (ii) Probabilistic Latent Semantic Analysis (PLSA), (iii) Latent Dirichlet Allocation (LDA), (iv) Correlated Topic Model (CTM).

"Latent Semantic Analysis (LSA) was first introduced in Dumais, Furnas, Landauer, and Deerwester (1988) and Deerwester, Dumais, Furnas, Landauer, and Harshman (1990) as a technique for improving information retrieval" [13]. LSA utilizes a matrix which its (i, j) entry indicates the occurrence of term i in document j. Then Singular Value Decomposition (SVD) is performed on the occurrence matrix to reconstruct the occurrence matrix into the multiply of three matrices.

Probabilistic Latent Semantic Analysis (PLSA) is introduced by Jan Puzicha and Thomas Hofmann that fixes some disadvantages in LSA. One big disadvantages of LSA is that it cannot distinguish polysemy and synonym. According to Alghamdi et al [4], "PLSA is a method that can automate document indexing which is based on a statistical latent class model for factor analysis of count data, and also this method tries to improve the Latent Semantic Analysis (LSA) in a probabilistic sense by using a generative model".

David Blei, Andrew Ng, and Michael Jordan came out with Latent Dirichlet Allocation (LDA) in 2003 [9]. Latent Dirichlet Allocation (LDA) is a model based on Bayesian topic models. In general, LDA tries to mimic writing process. The key idea behind LDA is that each document can be represented by a set of topics and each topic can also be represented by a set of words. LDA used Dirichlet distribution to calculate the probability of a word belongs to a topic.

Several researchers came up numerous of extensions of LDA to solve the limitation of LDA. A limitation of LDA is the inability to model topic correlation [8]. Correlated Topic Model (CTM) is introduced by David Blei who is also the researcher who developed LDA and John Lafferty in 2006 [8]. The key of the CTM is logistic normal distribution. In sum, CTM is a LDA that swaps the Dirichlet distribution with a logistic normal distribution.

2.3 Dynamic Networks

Visualization techniques for social networks include visually analyzing three types of networks: follower network, diffusion network, and reposting network. Follower Network encodes users as nodes and their follow-relation as the edges. Whereas, in the Diffusion Network, nodes represent messages while the edges represent a mention. Finally, the reposting Network combines the features of the first two networks by encoding users as the nodes and their reposting relation as the edges. [10]In our study, we are focusing on diffusion network visualization techniques because we want to investigate how different sentiments and topics related to COVID-19 vaccination propagate in the United States.

A lot of work is done in understanding diffusion networks and how information spreads in social networks. Google+Ripples [34] presents a diffusion network visualization technique that is a hybrid of node-link and circular treemap metaphors. Different messages have different sizes representing their diffusion rates. Since the reposting behavior builds a multi-level hierarchical structure, other studies focused on presenting a multi-layout view of diffusion networks. WeiboEvents [24] presents a visualization technique comprising of three layouts - a tree layout, a circular layout, and a sail layout. Extending on the sail layout, Li et al. [19] present a visualization technique that leverages parallel coordinates to represent the diffusion. One of the key motivations to study diffusion networks is to identify misinformation or anomalies. FluxFlow [37] is a system that enables users to visualize and identify the spread of anomalous information on social networks. Since we are interested in visualizing the diffusion network formed around the COVID-19 vaccination discussion on Twitter, we are combining FluxFlow and geospatial visualization techniques to understand how the American public's attitude around COVID-19 vaccines evolved.

Our project is going to focus on tweets about COVID-19 vaccination in the US. Other than developing new algorithms, our goal is to provide an interactive tool to users to do EDA and find patterns and insight. In addition, we are going to connect sentiment analysis, topic modeling, geospatial and temporal analysis instead of just focusing on one type of analysis like previously mentioned papers. As far as we could look, ours is the first paper that aims to explore and present the relationship between COVID-19 opinions expressed in tweets over time.

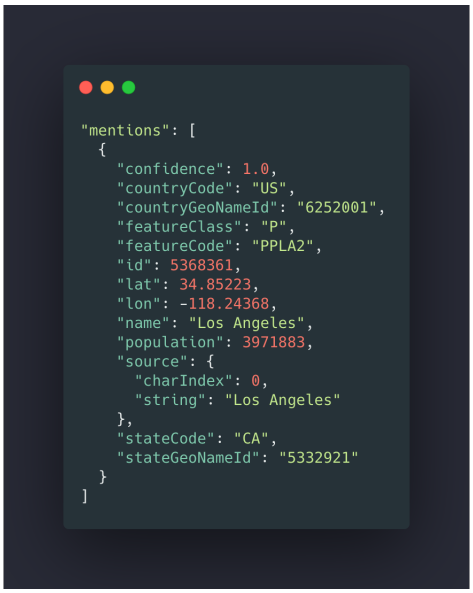

Fig. 1: Geoparsing Output

3 Proposed Solution

The driving motivation behind our design is to build a social media visual exploratory data analysis tool that combines textual, temporal, and geospatial visualizations. We divide our effort into three steps: data collection, data analysis, and data visualization. Data collection details our efforts in collecting the Twitter data relevant to COVID-19 vaccines. On the other hand, data analysis reflects the analysis decisions we took. Finally, data visualization describes our effort of combining different visualizations and the design guidelines we followed. We explain each step in more detail in the upcoming sub-sections.

3.1 Data Collection and Processing

Collecting social media data is tricky due to the private nature of the data. Our initial goal was to build a scalable and maintainable pipeline that automates Twitter data collection. In the beginning, we directed our efforts towards collecting the tweets directly from the Twitter API, hydrating tweets, filtering the relevant tweets, and storing the filtered data in a persistent database. However, we soon ran into technical issues from components beyond our control - Twitter API, external hydrating packages, etc. Our pipeline, albeit more scalable, was introducing delays in our project. Therefore, we decided to work with an open-sourced COVID-19 vaccine tweets dataset on Kaggle [23].

The Kaggle dataset, maintained by Gabriel Preda, contains tweets mentioning any of the following vaccines: Pfizer, Moderna, Sinopharm, Sinovac, AstraZeneca, Covaxin, and Sputnik V. It has 354,014 tweets from all parts of the world. Each record contains the following properties: *tweet id, username, user location, user description, user created, user followers, user friends, user favorites, user verified, tweet date, tweet text, tweet hashtags, tweet source, retweets, favorites, and is retweet*.

To extract the relevant tweets, we performed several transformations on the tweets. First off, we leverage geoparsing to keep the tweets that originated from within the United States. The geoparsing step returns the prediction of a city, state, country, latitude, and longitude given a string. The output for Los Angeles is shown in Fig 1.

Once we obtained the tweets from the United States, we performed several preprocessing steps on each tweet. Since we would be conducting sentiment analysis and topic modeling for each tweet, we needed to ensure that the tweets were in a format acceptable to the models. Our preprocessing pipeline consisted of the following immutable transformation steps: removing users from the tweets, removing links from the text, removing hashtags, removing audio and video links, converting text to lower case, stripping punctuation, removing double spacing, removing numbers, lemmatizing tweets, and finally tokenizing the tweets. A summary of data preprocessing steps is presented in Fig 2 At this stage, our dataset consisted of 17,275 tweets. Now we needed to extract some useful information from each tweet for creating the visualization.

3.2 Data Analysis

Since our goal is to enable users to explore public attitudes on COVID-19 vaccines, we extracted semantic information from each tweet. We focused on two sources of semantic information from text: sentiment analysis and topic modeling.

3.2.1 Sentiment Analysis

For sentiment analysis, we used the TextBlob package - a lexicon-based sentiment analysis toolkit. The data cleaning steps that we performed earlier allowed us to feed our data to sentiment analysis models after some additional processing. More specifically, we prepared the tweets by tokenizing them. Tokenization is the process of breaking down the text into a list of words. Furthermore, using TextBlob, we did the part-of-speech (POS) tagging. Here, we annotate each word with the appropriate part-of-speech tag: noun, verb, adjective, etc. After POS tagging, we conduct lemmatization, which is the process of replacing each word with its base word. As an example, the word "running" would become "run". Finally, we convert the text corpus into an N-gram model, where a combination of words is grouped to conserve context.

Once the prepared text is fed to the sentiment analysis model, it returns two properties: polarity and subjectivity. Polarity is a score between -1 and 1, which measures the negativity and positivity in text. A score of negative one denotes negative sentiment text, whereas a positive one denotes positive sentiment text. Subjectivity, on the other hand, refers to whether a text involves opinions or facts. A subjectivity score of zero indicates a highly subjective text, whereas a positive one represents a highly objective text. We performed sentiment analysis on each tweet and marked it negative, neutral, or positive based on the score. Any tweet between -1 and -0.33 was marked negative, between -0.33 and 0.33 as neutral, and between 0.33 and 1 as positive.

3.2.2 Topic Modeling

Our tool must explore what topics emerged as a part of public discourse on COVID-19 vaccines. For topic modeling, we performed the same preprocessing steps as we did for sentiment analysis. Topic modeling for tweets is challenging since each tweet is less than 140 characters long and usually contains only one topic. Therefore, LDA usually does not perform very well on tweets or any other short text. To mitigate the problems that arise with using LDA, we used a variant of LDA called the Gibbs Sampling Dirichlet Mixture Model (GSDMM). GSDMM assumes that each text (tweet) belongs to only one topic, whereas LDA assumes each text is a mixture of topics.

GSDMM model takes in a few hyperparameters. First off, we need to determine beforehand how many topics we wish to consider. Next, similar to LDA, we have two controlling hyperparameters: alpha and beta. Alpha determines how easily a topic gets removed when it's empty. As a result, the number of non-empty topics will get larger as alpha increases. On the other hand, beta controls how a topic is chosen for a tweet based on similarity to the topic rather than the popularity of the topic. As a result, the number of non-empty topic clusters will get smaller as beta increases. In short, a low alpha leads to fewer topics and a low beta leads to more similar topics.

3.3 Data Visualization

Based on the clean feature set we get as the output of the Data Preparation layer, we design several visualizations to highlight the diverging attitudes on COVID-19 vaccines in the United States. Our main goal is to provide a visual exploratory data analysis (EDA) tool for social media posts. Keeping our goal in mind, we deemed it important to connect different types of visualization in one view. From our data analysis, we were able to acquire two types of information: textual information and geospatial information. Through textual information, we aim to visualize how different sentiments and topics related to COVID-19

```
def preprocess_tweet(tweet):
    """Main master function to clean tweets, stripping noisy characters, and tokenizing use
lemmatization"""
    tweet = remove_users(tweet)
    tweet = remove_links(tweet)
    tweet = remove_hashtags(tweet)
    tweet = remove_av(tweet)
    tweet = tweet.lower()  # lower case
    tweet = re.sub('[' + punctuation + ']+', ' ', tweet)  # strip punctuation
    tweet = re.sub('\s+', ' ', tweet)  # remove double spacing
    tweet = re.sub('([0-9]+)', '', tweet)  # remove numbers
    tweet_token_list = tokenize(tweet)  # apply lemmatization and tokenization
    tweet = ' '.join(tweet_token_list)
    return tweet
```

Fig. 2: Preprocessing Steps

vaccines evolve over time. Geospatial information would help us map tweets to the US states and visualize how the attitudes diverge on a state-level.

In deciding a unifying view for the different types of visualizatios, we took inspiration from the NYC Foodiverse project by Will Su [31]. Our design criteria followed a few principles, namely:

1. **Preserve Context**: We tried incorporating sentiment, topical, geospatial, and meta information in one view. We believe it would help the users in their data analysis as they do not have to switch screens, preserving the visual context. Following this principle, we believe our tool improves the exploratory data analysis (EDA) process.

2. **Interactive Exploration**: Our second principle aims to enhance the information retrieval process for the users by enabling them to visually filter the tweet data. Since, we're presenting our tool as an EDA tool for social media, it is imperative that we allow users to interactively extract information.

3. **Detail View**: Finally, we tried to incorporate the meta data for each tweet and topic as we believe it will help the user build a narrative in their EDA process. We decided to hide the personal information such as the user name, user description, and user location in order to protect the user privacy. Even the geospatial view shows a view on the state level to preserve Twitter user identity.

To implement the visualization, we used D3.js on the front-end. On the backend, we used Python to collect, clean, analyse, and serve the data. We generously used open sourced tool at various steps in our pipeline: geoparsing packages to get location for each tweet, Pandas for data cleaning, TextBlob for sentiment analysis, and gsdmm for topic modeling. To integrate the geospatial view, we used the mapbox API and integrated its map view. For the color scheme, we used Dracula's color scheme [25].

4 Example Use Case/Initial Evaluation

Our interactive visualization webpage is an EDA (Exploratory Data Analysis) tool that enables users to freely explore the COVID-19 vaccination related Twitter dataset obtained from Kaggle. An example use case for our visualization can be a behavioral analyst who wants to understand the general attitude of the US population and then narrowing down the visualization to only display tweets from the state of California. To complete this use case, she will go through the following process:

Step 1: The analyst visits the webpage containing the COVIZVAX visualization. The first thing that shows up is the cover page which contains a short description of the visualization and the basic shape of the visualization. By clicking on the "Let's Explore" button, it leads her to an interactive layer allowing her to perform various user interactions (see Fig 3). The visualization wheel displays topics as words in the center and tweets as small circles around the circumference of the wheel. The colour of each circle represents the sentiment score obtained from each tweet: orange is negative (sentiment score: [-1, -0.33)), blue is neutral (sentiment score: [-0.33, 0.33)), and green is positive (sentiment score: [0.33, 1]).

Step 2: Now that the analyst is at the interactive page, she can zoom in and out, drag, and hover over the visualization wheel shown in Fig 3. Zooming in and out enables her to focus and unfocus on particular sections of the wheel. Additionally, the zoom function can help her achieve precision but also to understand the entire picture. The dragging function enables her to move the wheel to different locations on the webpage. When hovering over a topic node, the side channel shows the description of the topic, a set of words associated with the topic, and the total number of Twitter posts related to this topic (see Fig 5). Hovering over a Tweet node highlights the node, the topic node connected to it, and the edge between the two. The side channel displays the text content and a collection of metadata about the Tweet containing the number of retweets, number of likes, location, creation date, topic description, and a map showing the latitude and longitude for the identified location. The map in the side channel is also zoomable, enabling her to view location data at various granularity (see Fig 4). All of these basic functions along with the side channel can help the analyst to obtain a complete picture of the general opinions in the US.

Step 3: To take look at the divergence of opinions of a specific state, the analyst can add or remove four different filters located in the navigation bar. The filters ranges by sentiment, state, topic, and date range. To apply the filters for this example use case, the analyst can select California from the drop down list. The selection will immediately update the visualization to exclude tweets that are not located in California. In addition, for examining the divergence of opinion, she can change the date change, and hover over the topic and tweet nodes to examine common patterns amongst the tweets that are in this location between the specified time range. If she wants to take a look at only positive tweets or tweets that belongs to a certain topic, this can be accomplished with the filters as well.

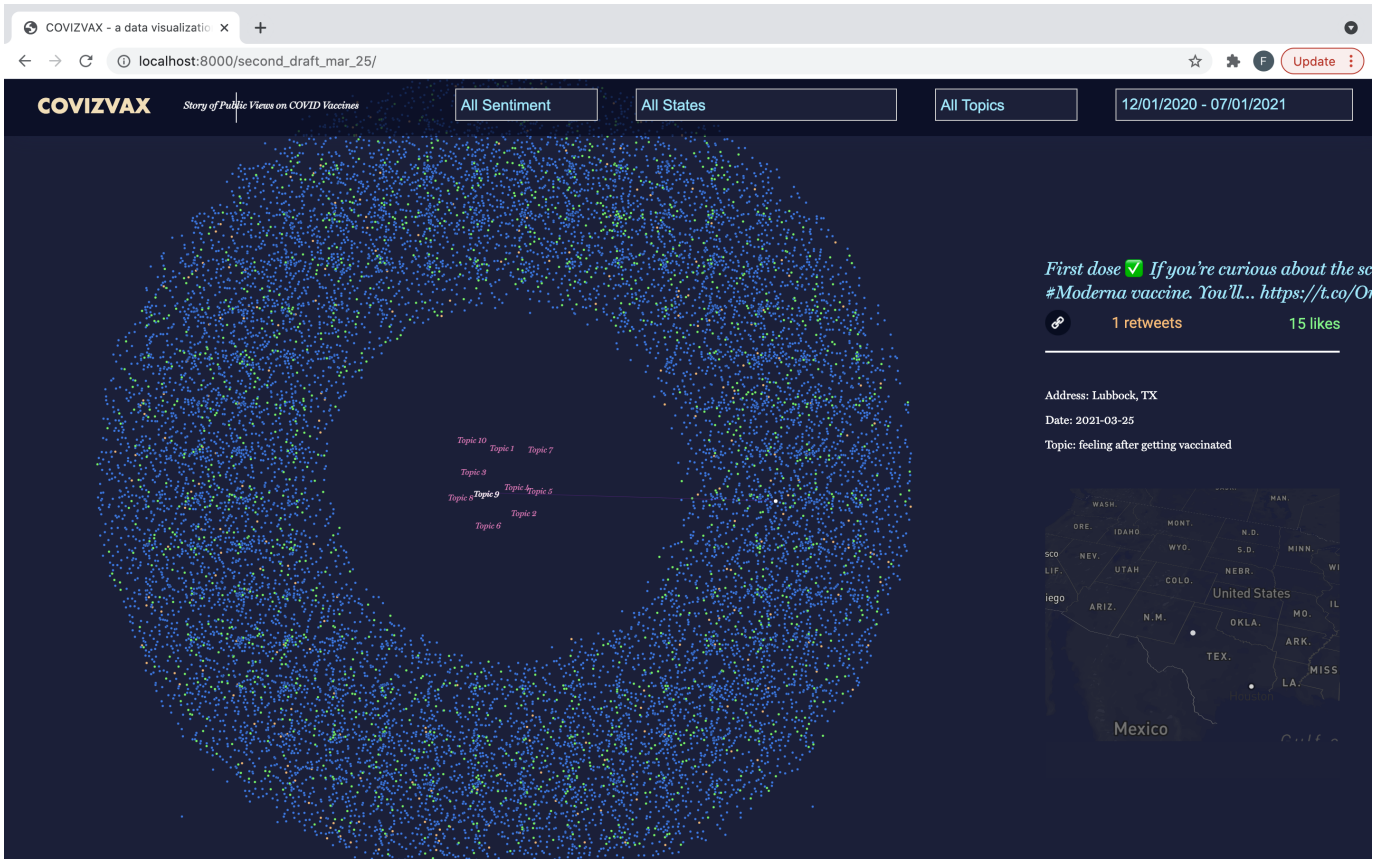

Fig. 3: Interactive Dashboard

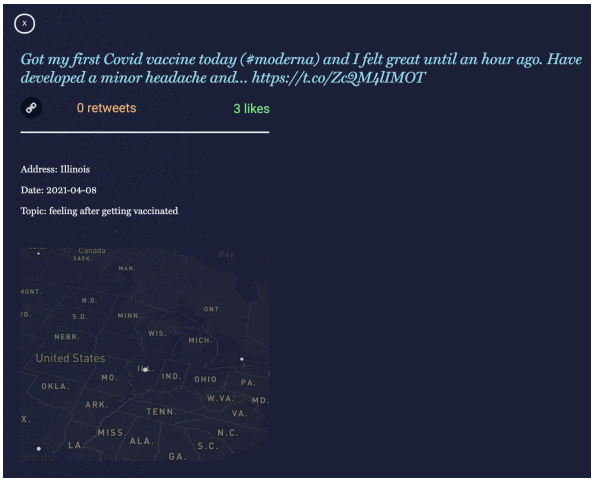

Fig. 4: Tweet Nodes - Expanded Side Channel

Fig. 5: Topic Nodes - Side Channel

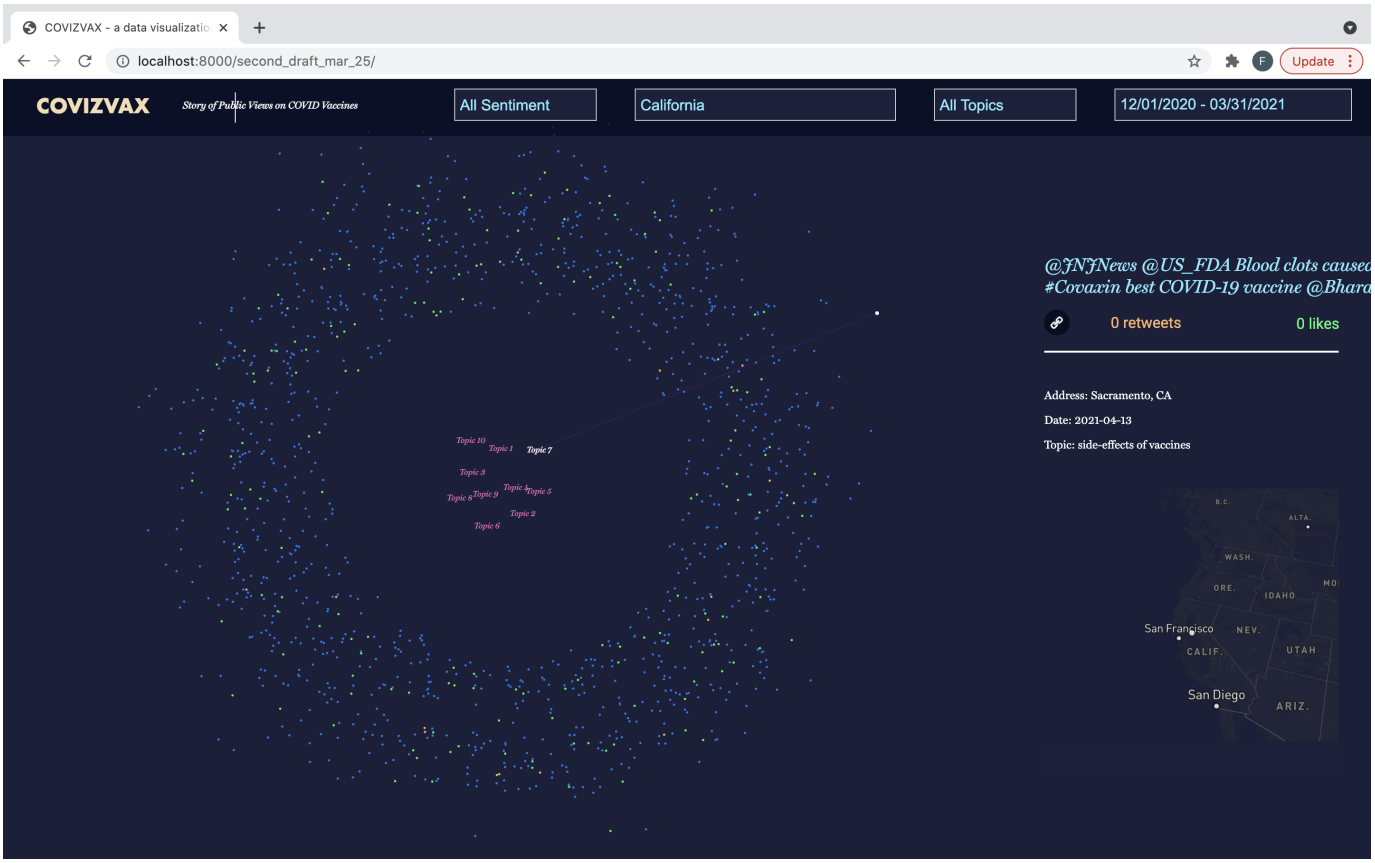

Fig. 6: Visualization with California and Date Range Filter

5 Discussion

This paper is part of our Information Visualization course requirement at the University of Waterloo. Being part of the course presented us with unique challenges and opportunities. We had to modify the project on a couple of occasions to fit the timeline of the course, and at the same time we were learning new techniques every week, some of which we were able to add to our project.

Our final visualization tool provides a great way for users to perform exploratory data analysis. It combines elements of geospatial analysis, sentiment analysis, temporal analysis, and topic analysis in an interactive manner. The tool is a direct result of our second research question: "Can a visual tool be built to provide interactive overview of the views on vaccines?". It also helps answer our first research question: "How are the views on COVID-19 vaccines distributed across the US?". Using the tool, the views on vaccines can be explored across any state in the US, for any selected time window. The filtering option further provides option to focus only on positive, negative, or neutral tweets.

Our visualization template builds on the work done in nycfoodiverse [31] and adds further dynamic filtering capabilities to it. We believe the resulting visualization to be a powerful tool that can be used by other researchers aiming to do EDA on social media data.

One challenge that we faced was with data collection. We first had to spend time to arrange for Twitter developer license to start collecting data. We then realized that Twitter API limits the data collection rate, and this caused some delays in collecting a sizeable amount of data that we could work on. Eventually, we had to work with a Twitter dataset available on Kaggle to save time.

Another challenge was, ironically, due to the very same reason that inspired our project topic: COVID-19. All the authors of this paper were working from different locations and were not able to be in the same room and work together. We instead had to rely on video conferencing and collaborating remotely, which, fittingly, reflects the zeitgeist of the current period.

We also made changes to our dashboard. We initially envisioned it to a collection of dynamically linked charts, each of which would focus on one aspect out of geospatial, topic, sentiment, etc. However, through our various lectures and readings, we were able to find a way to combine some of those elements into a single powerful interactive visualization. This design change was a direct result of our increased capabilities as visualization specialists, as we honed our skills while working on the Visual Design exercises.

Finally, there are also some limitations of the project. We had initially hoped to incorporate some sort of network analysis of the Twitter dataset, but had to leave it out due to time constraints.

6 Future Work

Despite being highly intuitive, COVIZVAX is limited to preprocessed COVID-19 data. However, the procedure of processing and visualizing textual data from social apps is extremely similar. Therefore, we plan to extend COVIZVAX to be able to visualize any raw textual data by integrating data extraction and processing steps into the tool. On the other hand, COVIZVAX currently uses a lexicon based method in sentiment analysis step. One disadvantage of this is that the sentiment score is solely based on the lexicon, hence the accuracy is not perfect in some circumstances. We would like to give users the freedom to customize the methods used in our data processing module instead of setting them up beforehand by developers. In this way, users will be able to process the data using their pre-trained or preferred machine learning module to get more accurate results.

We would like to broaden the type of charts in the interface of COVIZVAX as well. So far, COVIZVAX only has a graph that links tweets with similar topic, and the location on the map if the user selects a particular tweet (node). An overview of locations of tweets on a map and a histogram of sentiment scores would be helpful for users to get insights from their data.

Thirdly, to further understand the relationship between tweets with

similar sentiment score or topic keywords, we want to integrate social network into the visualization. Our goal here is to form a tweet network and apply community detection algorithms to understand what factor could affect people's view on vaccination. For instance, are the views on vaccination driven by some subtle underlying network patterns? Do politics and politicians' perspective change the public's view on vaccination?

Lastly, the performance of COVIZVAX is not ideal when the visualization contains too many data points (tweets). Graph generation and interactions are laggy due to the number of data points. This is a common known issue with D3. We plan to improve the performance in order to generate smooth interactions.

7 Conclusion

As the vaccination rates have become stagnant, and COVID-19 variants like delta variant continue to claim lives and hospitalize people, it is becoming more and more important to understand people's views on vaccination. Understanding the main topics that people talk about on social media platforms, what the sentiments of these people are is an important step in addressing vaccine hesitancy and misinformation. It is also important to understand how these sentiments geospatially and with time.

To accomplish this, our visual dashboard can be a good tool that can help anyone looking to understand these trends. The interactive and visual nature of the dashboard helps the users perform exploratory data analysis efficiently.

Our tool provides a visualization framework that can be applied to other problems of public interest, where understanding people's sentiments and important topics is important. To conclude, we believe that our tool could be useful in situations that demand careful monitoring of public attitudes.

Acknowledgments

Below are the main contributions of our teammates:

- **Fa Fa Ke**: Introduction, Related Work, Use Case/Initial Evaluation, Geoparsing, Video Demo, Visualization Tool Filters.
- **Abdullah Mobeen**: Introduction, Related Work, Proposed Design, Data Formatting, Topic Modeling, Visualization Tool.
- **Abhishek Sharma**: Abstract, COVID-19 key research areas, Discussion, Conclusion, geospatial analysis, Visual interaction editing, general editing, spelling and grammar
- **Mushi Wang**: Abstract, Introduction, Related Works, Sentiment Analysis, Future Works, general LaTeX editing.

References

[1] A. Abbasi and H.-c. Chen. Cybergate: A design framework and system for text analysis of computer-mediated communication. *MIS Quarterly*, 32(4), 2008.

[2] W. Ahmed, J. Vidal-Alaball, F. Lopez Segui, and P. A. Moreno-Sánchez. A social network analysis of tweets related to masks during the covid-19 pandemic. *International Journal of Environmental Research and Public Health*, 17(21):8235, 2020.

[3] M. Al-Ramahi, A. Elnoshokaty, O. El-Gayar, T. Nasralah, and A. Wahbeh. Public discourse against masks in the covid-19 era: Infodemiology study of twitter data. *JMIR Public Health and Surveillance*, 7(4):e26780, 2021.

[4] R. Alghamdi and K. Alfalqi. A survey of topic modeling in text mining. *Int. J. Adv. Comput. Sci. Appl.(IJACSA)*, 6(1), 2015.

[5] S. Baccianella, A. Esuli, and F. Sebastiani. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec*, vol. 10, pp. 2200–2204, 2010.

[6] L. Barbosa and J. Feng. Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, p. 36–44. Association for Computational Linguistics, USA, 2010.

[7] A. Bifet and E. Frank. Sentiment knowledge discovery in twitter streaming data. pp. 1–15, 09 2010.

[8] D. Blei and J. Lafferty. Correlated topic models. *Advances in neural information processing systems*, 18:147, 2006.

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[10] S. Chen, L. Lin, and X. Yan. Social media visual analytics. *Computer Graphics Forum*, 36(3), 2017.

[11] M. Cinelli, W. Quattrociocchi, A. Galeazzi, C. M. Valensise, E. Brugnoli, A. L. Schmidt, P. Zola, F. Zollo, and A. Scala. The covid-19 social media infodemic. *Scientific Reports*, 10(1):1–10, 2020.

[12] M. R. DeVerna, F. Pierri, B. T. Truong, J. Bollenbacher, D. Axelrod, N. Loynes, C. Torres-Lugo, K.-C. Yang, F. Menczer, and J. Bryden. Covaxxy: A collection of english-language twitter posts about covid-19 vaccines. *arXiv preprint arXiv:2101.07694*, 2021.

[13] S. T. Dumais. Latent semantic analysis. *Annual review of information science and technology*, 38(1):188–230, 2004.

[14] S. Govind. Analyzing twitter trends using ai, python. https://medium.com/analytics-vidhya/analyzing-twitter-trends-using-ai-python-e0aee72fde87, 2020.

[15] R. Hernandez, L. Hagen, K. Walker, H. O'Leary, and C. Lengacher. The covid-19 vaccine social media infodemic: Healthcare provider's missed dose in addressing misinformation and vaccine hesitancy. *Human Vaccines & Immunotherapeutics*, 17, 04 2021. doi: 10.1080/216445515.2021.1912551

[16] A. Hussain, A. Tahir, Z. Hussain, Z. Sheikh, M. Gogate, K. Dashtipour, A. Ali, and A. Sheikh. Artificial intelligence–enabled analysis of public attitudes on facebook and twitter toward covid-19 vaccines in the united kingdom and the united states: Observational study. *Journal of medical Internet research*, 23(4):e26627, 2021.

[17] S. N. Keller, J. C. Honea, and R. Ollivant. How social media comments inform the promotion of mask-wearing and other covid-19 prevention strategies. *International Journal of Environmental Research and Public Health*, 18(11):5624, 2021.

[18] K. Kucher, C. Paradis, and A. Kerren. The state of the art in sentiment visualization. In *Computer Graphics Forum*, vol. 37, pp. 71–96. Wiley Online Library, 2018.

[19] Q. Li, H. Qu, L. Chen, R. Wang, J. Yong, and D. Si. Visual analysis of retweeting propagation network in a microblogging platform. In *Proceedings of the 6th international symposium on visual information communication and interaction*, pp. 44–53, 2013.

[20] A. Lovari. Spreading (dis)trust: Covid-19 misinformation and government intervention in italy. *Media and Communication*, 8:458, 06 2020. doi: 10.17645/mac.v8i2.3219

[21] M. Müller, M. Salathé, and P. Kummervold. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. arxiv 2020. *arXiv preprint arXiv:2005.07503*.

[22] W. H. Organization. Tracking sars-cov-2 variants. https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/, 2021.

[23] G. Preda. Covid-19 all vaccines tweets. Kaggle.com, August 2021.

[24] D. Ren, X. Zhang, Z. Wang, J. Li, and X. Yuan. Weiboevents: A crowd

sourcing weibo visual analytic system. In *2014 IEEE Pacific Visualization Symposium*, pp. 330–334. IEEE, 2014.

[25] Z. Rocha. Dracula color theme. Dracula, August 2021.

[26] S. Rufai and C. Bunce. World leaders' usage of twitter in response to the covid-19 pandemic: a content analysis. *Journal of public health (Oxford, England)*, 42, 04 2020.

[27] H. Saif, Y. He, M. Fernandez, and H. Alani. Contextual semantics for sentiment analysis of twitter. *Information Processing & Management*, 52(1):5–19, 2016.

[28] D. Schillinger, D. Chittamuru, and A. S. Ramírez. From "infodemics" to health promotion: A novel framework for the role of social media in public health. *American Journal of Public Health*, 100(9), 2020.

[29] K. Sengul. Never let a good crisis go to waste: Pauline hanson's exploitation of covid-19 on facebook. *Media International Australia*, 178:1329878X2095352, 08 2020. doi: 10.1177/1329878X20953521

[30] D. Stojanovski, I. Dimitrovski, and G. Madjarov. Tweetviz: Twitter data visualization. In *Conference on Data Mining and Data Warehouses (SiKDD 2014) - Information Society*, 2014.

[31] W. Su. Nyc foodiverse. nycfoodiverse.com, August 2021.

[32] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267—-307, 2011.

[33] S. G. Tan, A. Sesagiri Raamkumar, and H. L. Wee. Users' beliefs toward physical distancing in facebook pages of public health authorities during covid-19 pandemic in early 2020. *Health Education & Behavior*, p. 10901981211014428, 2021.

[34] F. Viégas, M. Wattenberg, J. Hebert, G. Borggaard, A. Cichowlas, J. Feinberg, J. Orwant, and C. Wren. Google+ ripples: A native visualization of information flow. In *Proceedings of the 22nd international conference on World Wide Web*, pp. 1389–1398, 2013.

[35] Y. Wang, Y. Kawai, K. Sumiya, and Y. Ishikawa. Tweevist: A geo-tweet visualization system for web based on spatio-temporal events. In *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, pp. 1–6, 2016. doi: 10.1109/ICIS.2016.7550845

[36] Worldometer. Covid-19 coronavirus pandemic. https://www.worldometers.info/coronavirus/, 2021.

[37] J. Zhao, N. Cao, Z. Wen, Y. Song, Y.-R. Lin, and C. Collins. # fluxflow: Visual analysis of anomalous information spreading on social media. *IEEE transactions on visualization and computer graphics*, 20(12):1773–1782, 2014.