Visualizing Covid-19 Sentiment and Factors that Affect It

Mohan Dodda
mdodda3@gatech.edu

Shaan Gill
sgill36@gatech.edu

Bori Han-Yang
bhanyang3@gatech.edu

Darryl Jacob
djacob30@gatech.edu

Trevor McCrary
tmccrary9@gatech.edu

Logan Schick
lschick3@gatech.edu

1 Introduction

We plan to visualize Covid-19 perception geographically over time. To do this, we look at sentiment and emotions of tweets, including positive and negative sentiment and emotions of anger, anticipation, disgust, fear, joy, negative, positive, sadness, surprise, and trust over time, and in each state. Additionally, we will look at the factors that affected Covid-19 Sentiment such as Vaccine Rates, Government Regulation, Political Leaning, and Impact of Popular Influencers using correlation experiments.

This Project will not cost anything other than electricity cost associated with coding and retrieving our data using our laptops.

Why does anyone want to know about changing COVID-19 sentiment? One of the primary challenges faced by public health officials during the COVID-19 pandemic was the management of public sentiment given the environment of widespread misinformation. By understanding why public perceptions swing during pandemics, we are enabling the future public health leaders to better manage future crises

2 Problem Definition

In the past few years, Covid-19 has been prolonged and people have gone through a locked-down society. Things that had been taken for granted in the past have collapsed, and in the process of normalization again, people's feelings about the epidemic had changed a lot to date than when the initial pandemic occurred. This is considered to have a lot to do with not only the epidemic itself, but also the social policies that control it, and the reactions of those who follow it. People's emotions, as mentioned above, we will use the Twitter API to collect data about people's tweets about COVID-19. Through this, we plan to conduct multi-level sentiment analysis according to time through our pre-designed emotion model. Research regarding this was only concerned with people's emotions per se, but this will vary from state to state. We believe that there is a correlation between changes in sentiment over time and variables such as political leaning, Vaccine Rates, and Government Regulation and analyze these factors.

3 Survey

Our main method of gathering data for this project is sentiment analysis of tweets. We reviewed a variety of papers about sentiment analysis and the challenges of analyzing tweets. We also looked into papers regarding Covid-19 sentiment specifically.

Sentiment analysis has various challenges due to polarity shifts, data sparsity, or inclusions of URLs. It's important to understand potential problems with our method and dataset to avoid issues if we can't extract information due to very complex language. [1] [2] [3]

There are algorithms that perform sentiment analysis on real time tweets related to Covid-19 to identify anomalous events and clusters tweets by keywords. While we are not studying real time data, our dataset spans multiple years[3] [4].

[5] In here Covid Perception is measured using the emotions with Twitter data which we described above. We'll follow a similar strategy and perform the factor analysis as well. We have also found cases where Naive Bayes is used which is more modern, but has the same issue of only showing sentiment[6]. We are attempting to get deeper insight by correlating with the events happening at the time of the tweets.

[7] In this one, general Covid perception is measured survey style and they find that level of education and partisanship and not race/ethnicity are factors that are associated with Covid-19 perception and vaccine hesitancy so we use partisanship as a factor for our project. This is how Covid Perception (Sentiment) and associated factors are mea-

sured nowadays but this survey style can only retrieve limited number of data and might not be representative of the general population.

[8] This paper proposes the general neural architecture of a model we plan to use for sentiment analysis and emotional analysis training. This paper introduces an encoder-decoder architecture using attention network [9] which allows the model to focus over specific part of the sentence during model training. [10] The transformer variant we are going to use is BERT the encoder part of it that is bidirectional reading sentences both ways. For both transformers and BERT, shortcomings are that they are really big and take a long time to train so we will do fine-tuning.

These papers talk about the times that one should avoid using the Pearson correlation coefficient and the potential problems with using it. One refers to when the data is not linearly suitable [11] and the other discusses how it performs poorly with outliers [12]. This is relevant to our project as we may need to adapt to our data and use a different correlation. These papers provide insights into when this may be necessary and the limitations of the Pearson coefficient. The limitations of both papers are that there is no definite way of knowing that the Pearson coefficient should be avoided and more provides general guidance which should be applied as needed.

Shellack et al. studied public sentiment on effectiveness of repurposed medicines like Hydroxychloroquine on Reddit and Twitter [13]. This paper really helps us because they also look at types of user accounts and their role in the spread of misinformation. For example, they identify politicians and healthcare worker accounts as particularly effective at spreading information. Shellack et al. does not model the differences in communication style across platforms and reduces sentiment signal. Our project scopes our data source to Twitter.

[14] This paper focuses on sentiment analysis directly related to the Covid-19 vaccine. It found that sentiment related to a vaccine's country of origin as well as daily news contributed to sentiment.

4 Proposed Method

Our goal is to better understand the connections between sentiment towards Covid-19 and how dealing with the pandemic was approached in each State. If we are able to gain insight into these connections then we would be able to identify what actions lead to positive changes such as increased vaccination rate. We can check the success of our method by finding positive correlation with our expectations of how people's sentiment towards Covid-19 affect their actions. To measure this we will be using the Pearson Correlation Coefficient. Finding these factors associated with emotion-specific sentiment is what is new in our approach compared to just regular sentiment. Our approach will be successful because all results will provide more information about the question "How does perception change?". For instance, if we find popular influencers do not affect sentiment, we now understand how to better inform the public. Our approach might fail and be risky in that we might not have enough data as the Twitter data might be spotty. Additionally, correlations of factors may be noisy and provide an imprecise signal-changing perception. Payoffs of our method will be that the gathered sentiment will be highly correlated with vaccine rate so people such as government officials can make decisions that maximise those sentiments if a similar situation occurs in the future.

We will be gathering our sentiment data from a dataset of Covid-19 related tweets [15]. We can gather the text from these tweets using Twitter's developer API and the Tweepy Python library, we will filter the data by only including tweets that had a location tag. The potential issue with collecting this data is there are rate limits for looking up tweets, but since we have already started and have gotten access to an education developer account with higher rate limits this shouldn't be a problem. Additionally, we can batch-request tweets to help alleviate the data caps that Twitter has as you can batch 100 tweets into 1 request. Once gathered we will measure the sentiment using the pre-trained NLTK sentiment analysis. We will also gather emotional data using the text2emotion library since it is easy to use and quick to run.

The gathered information will be cleaned and combined using pandas dataframes. Pandas is the optimal choice for this because it has functionality for dealing with dates, which will help us line up the data sets, and also has good grouping functionality so metrics like average sentiment per state, per week, and per month can be easily calculated.

Our primary innovation is to connect the classified sentiment data we gather to three axes that we have not seen any previous papers examining. The axis that we believe should have the most statistical significance is stated vaccine rates. In conjunction with the vaccine rates, we can look at the lockdown regulations used in each state for connections between the strictness of regulations, sentiment in the state, and vaccine rate. The vaccine rate and regulation data we get from government websites such as the CDC. By examining this data over time, by week and by month, we can look for inflection points to see if a change in people's perception of COVID leads to an increase in the number of people getting vaccinated. We can also look at what tweets are most popular to see what information may have had the largest impact.

To make it easier to interpret our results we will be creating a website to visualize the data we generated. This website will include bar graphs to show the levels of sentiment in tweets at a particular time, choropleths to show various datapoints such as vaccine rate and emotional analysis in each state in a particular month, and scatterplots to show trends in any pair of axis. To make these visualizations we used D3 because it is a robust library with the ability to make the plots we desire.

5 Experiments

We plan on retrieving the Twitter dataset that has Covid-19 related tweets for each state. We want enough data over different time periods so we can see the tweet sentiment over time. We utilize the Twitter API to collect all our tweets with the state-divided Tweet IDs. To do this we needed to batch our requests to speed up the process due to data caps on twitter's API. Additionally, we need to request all of the relevant fields that we want and also format the twitter response to receive all of these relevant fields. Once the data is been processed the locations need to be further processed since the Twitter API does not directly give the location in batch requests. Once we have retrieved all of the data we can put it into our own CSV file for further processing. We had about 650,000 tweets that we pulled however we needed to remove all tweets that returned an error (were deleted by user or Twitter) which resulted in a dataset with about 550,000. In terms of fields retrieved, this included: Tweet id, author id, location (full name/name), created at, retweet count, reply count, like count, quote count, source, language, is sensitive, and bounding box. Due to the format of the Twitter API, we also need to further format the location to extract only the state, and Twitter formats it differently depending on what the settings the user has set.

After we get the relevant data we need to calculate the appropriate sentiments. We set this experiment by utilizing pretrained NLTK sentiment analysis and the pretrained text2emotion. We attempted to run a transformer model that was available on hugging face however due to the compute time and the results it was giving we decided to go with the text2emotion library. The transformer model was taking about 10 times longer than the other model and due to our large database, we did not think it was a good idea to use this. Additionally, some tweets on Twitter were simply a hyperlink and the transformer model would assign a value to this while the Python package would place all the emotion weights at 0. Additionally, the Python package gives scores for all the emotions while the transformer model only returned the highest score. Due to these differences, we decided to go with the text2emotion Python package. We run these on all of our tweets evenly to get positive/negative sentiment score and get the emotion scores. We will also retrieve the data on vaccine rates, government regulation, and political leaning. For vaccine rates, we have a dataset that shows vaccinations per state at every month. For political leaning, we have datasets that show statewide election results nationally using 2020 Presidential results. The election results are not granular time-wise so we can mainly do spatial analysis. Lastly for government regulation, we will look at statewide reopening and regulations and when the timing of them.

Now for each state and nationwide in general, we do correlation experiments. The first one is with Covid sentiment/emotions with vaccine rates. We see if there is positive or negative correlation between these two using Coefficient of Determination $R\hat{2}$ over time for each state. We will also look to see if there are any spikes in vaccine rates and if big government regulations correspond with spikes in certain tweet sentiments/emotions. We will see if there sentiment precedes or succeeds the vaccine rates or big government regulation. With a high enough Pearson Coefficient (P-value), we can hypothesize that negative or positive sentiment or certain emotions caused government regulation and/or increase of taking vaccines or if certain sentiment were caused by government regulation.

We analyzed the entirety of our dataset to find the most influential users. We judged this by finding the users with the most tweets related to Covid-19

Correlation	Neg	Pos
Dose1	0.014237	-0.098498
Dose2	0.028950	-0.100877

Table 1: Pearson Correlation Coefficient calculation between first and second dose population percentage and negative and positive sentiment

Correlation	Happy	Angry	Surprise	Sad	Fear
Dose1	-0.067132	0.012397	-0.043897	0.059292	0.071491
Dose2	-0.072087	0.012571	-0.048422	0.062360	0.072517

Table 2: Pearson Correlation Coefficient calculation between first and second dose population percentage and all emotion values

as well as the users with the most likes on tweets related to Covid-19. The result of this experiment revealed no overlap between the top 20 users in both categories. When expanding to the top 50 users in both categories, there was only 1 user that was in both categories. This user is a law professor at UC Hastings. When analyzing the top 10 users in categories there is a large trend differentiating between the two groups. The users with the most tweets tended to have a more negative sentiment score and tended to be ordinary people, whereas the users with the most likes tended to be individuals in the medical field or individuals involved in news media. The users with the most liked tweets also tended to have a close to neutral sentiment score in their tweets.

When examining the average emotional value of tweets in each state for each month we found that states were consistently exhibiting fear as the highest emotional value. Since the level of fear is so consistent it is difficult to draw conclusions about the relationship between fear and vaccine rate. For example in April 2021, the third month of the Covid vaccine being wildly available, Connecticut had one of the highest first dose vaccine rate at 46.237% and an average level of fear in tweets of 0.24, while one of lowest, Alabama, had a first dose rate of 28.997% and fear of 0.229. Despite this, our calculations do show a positive correlation of fear and vaccine rate in both doses, which can be exemplified by New Hampshire which consistently had fear above 0.3 and had the highest vaccine rate from month to month.

While we weren't able to find major differences in sentiment and emotional values the correlations show what one would expect when predicting how sentiment or emotion would impact the vaccine rate. People that have negative sentiment, and are sad or fearful should be more likely to get vaccinated to gain some measure of safety from the epidemic. People who are happy are less likely to be phased by the epidemic and instead believe they'll be fine without the vaccine. It was also noticeable that in general states that leaned Democrat had higher vaccine rates than those that leaned Republican. We also cross-referenced dates when government regulations changed, such as the removal of restrictions on sitting in restaurants in a particular state, to see if there was a significant change in sentiment in that or the following months, but there were no significant changes.

6 Conclusion

This project focused on the relationships between Twitter sentiment and vaccination rates, government policy, and political leaning. We accomplished this by gathering data from the Twitter API and inferring sentiment with political leaning. To examine the effects of government policy, we are also tabulating 2020 election results by state and examining their influence on Twitter sentiment. Finally, we visualized this data using a plethora of techniques from choropleth maps of Twitter sentiment against vaccination rates to bar charts of Twitter sentiment by the length of time the account has existed.

One limitation of our project was the reliance on tweets with location data for our analysis. This creates a small bias in our data for people that enabled this feature.

The implication of this project is to show an idea of the general public's sentiment toward the Covid-19 pandemic throughout the course of the vaccine rollout. We analyzed our dataset to discover the most influential users based on the amount of tweets and amount of likes. These two groups had little crossover between them. The Covid-19 related tweets with the most likes tended to be from reliable sources such as medical professionals and people in news media. This shows a strong resistance for the likelihood of misinformation receiving a large amount of engagement. The only major connection we were able to find between the low and high vaccination rate states was their political leaning, but our correlation calculations were able to confirm some relationship between the sentiment and emotions of the tweets in that negative sentiment emotions tended to be positively correlated with vaccine rate and the opposite for positive sentiment and emotions.

This project could be expanded upon by gathering more Twitter data and broadening its scope. While our project focused on tweets in English with location data, a more detailed picture of global sentiment could be obtained by looking at Twitter data from other languages. It is possible that looking at this data from a global scale since people were affected by Covid on a global scale as opposed to a state scale. It is also possible to analyze the data with other variables such as infection rate or severity of government lockdown policies.

All team members have contributed a similar amount of work.

References

[1] A. M. Abirami and V. Gayathri. A survey on sentiment analysis methods and approach. In *2016 Eighth International Conference on Advanced Computing (ICoAC)*, pages 72–76, 2017.

[2] Monireh Ebrahimi, Amir Hossein Yazdavar, and Amit Sheth. Challenges of sentiment analysis for dynamic events. *IEEE Intelligent Systems*, 32(5):70–75, 2017.

[3] Guillermo Blanco and Anália Lourenço. Optimism and pessimism analysis using deep learning on covid-19 related twitter conversations. *Information Processing Management*, 59(3):102918, 2022.

[4] Bakhtiar Amen, Syahirul Faiz, and Thanh-Toan Do. Big data directed acyclic graph model for real-time covid-19 twitter stream detection. *Pattern Recognition*, 123:108404, 2022.

[5] Sakun Boon-Itt and Yukolpat Skunkan. Public perception of the covid-19 pandemic on twitter: Sentiment analysis and topic modeling study. *JMIR Public Health Surveill*, 6(4):e21978, Nov 2020.

[6] Kamaran H Manguri, Rebaz N Ramadhan, and Pshko R Mohammed Amin. Twitter sentiment analysis on worldwide covid-19 outbreaks. *Kurdistan Journal of Applied Research*, pages 54–65, 2020.

[7] Vitalis C Osuji, Eric M Galante, David Mischoulon, James E Slaven, and Gerardo Maupome. Covid-19 vaccine: A 2021 analysis of perceptions on vaccine safety and promise in a us sample. *Plos one*, 17(5):e0268784, 2022.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[9] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. *CoRR*, abs/1508.04025, 2015.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[11] Richard A Armstrong. Should pearson's correlation coefficient be avoided? *Ophthalmic and Physiological Optics*, 39(5):316–327, 2019.

[12] Yunmi Kim, Tae-Hwan Kim, and Tolga Ergün. The instability of the pearson correlation coefficient in the presence of coincidental outliers. *Finance Research Letters*, 13:243–257, 2015.

[13] Natalie Schellack, Morné Strydom, Michael S. Pepper, Candice L. Herd, Candice Laverne Hendricks, Elmien Bronkhorst, Johanna C. Meyer, Neelaveni Padayachee, Varsha Bangalee, Ilse Truter, Andrea Antonio Ellero, Thulisa Myaka, Elysha Naidoo, and Brian Godman. Social media and COVID-19—perceptions and public deceptions of ivermectin, colchicine and hydroxychloroquine: Lessons for future pandemics. *Antibiotics*, 11(4):445, March 2022.

[14] Han Xu, Ruixin Liu, Ziling Luo, and Minghua Xu. Covid-19 vaccine sensing: Sentiment analysis and subject distillation from twitter data. *Telematics and Informatics Reports*, 8:100016, 2022.

[15] Shay Palachy. awesome-twitter-data, 2022.