Using Reddit Comment Sentiment Analysis as a Proxy for COVID-19 Vaccine Acceptance by State.

JAKE CUPANI, University of Maryland, USA

1 ABSTRACT

Sentiment analysis is used extensively in the context of social media in an attempt to understand the public's opinion on various topics. With the unprecedented spread of the COVID-19 pandemic, usage of social media as a forum for discussion on COVID-19 vaccines has been nothing short of divisive. Reddit is one popular social media platform in which these discussions commonly take place. Comments from these COVID-19 and vaccine related posts have been given a polarity score from -1 to 1 to measure negativity and positivity of top-level comments, respectively. Data on vaccination rates by state is then used to correlate with these polarity scores. Three different techniques are used to obtain polarity scores for each top-level comment. These techniques are TextBlob basic polarity scoring, title and comment matching TextBlob polarity scoring, and GPT-3 polarity scoring. While GPT-3 polarity scoring and basic TextBlob polarity scoring exhibited moderate positive correlation with average vaccination rates by state, the title and comment TextBlob polarity matching method was not as successful. Nonetheless, this study shows how sentiment analysis metrics such as polarity can be used as a proxy for vaccine acceptance by state in the United States.

Additional Key Words and Phrases: COVID-19; GPT-3; TextBlob; public sentiment; social media; Reddit; Misinformation; communication; sentiment analysis; pandemic; vaccine; hesitancy; United States;

ACM Reference Format:
Jake Cupani. 2022. Using Reddit Comment Sentiment Analysis as a Proxy for COVID-19 Vaccine Acceptance by State.. *ACM Trans. Graph.* 37, 4, Article 111 (August 2022), 5 pages. https://doi.org/XXXXXXX.XXXXXXX

2 INTRODUCTION

The SARS-CoV-2 virus, or COVID-19 virus, which is a new type of coronavirus discovered in 2019, has created a global pandemic that has fundamentally changed many aspects of society and public health. In fact, at the time of this paper, there have been approximately 995,115 total deaths from COVID-19 [15]. Alongside the ongoing health crisis, vaccines have been developed by healthcare distributors such as Pfizer, Moderna, and Johnson & Johnson to combat the spread of the COVID-19 virus. As of March 2022, over 258 million people have at least one dose of the COVID-19 vaccine [5]. While the COVID-19 vaccine has proven to be both safe and effective at preventing severe illness from the virus [1], it has received varying levels of hesitancy from the general public [2]. This poses an inherent challenge to controlling the spread of the virus. There are many factors that contribute to vaccine acceptance. Loomba et al. found that exposure to misinformation subsequently lowers one's intent in accepting the COVID-19 vaccine [4]. Alamoodi et al. describes that factors such as misinformation, distrust in healthcare systems, and lack of understanding are major contributors to vaccine hesitancy and vaccine uptake [14]. Furthermore, in order for vaccines to be effective, a high uptake must be achieved [9], thus emphasizing the importance of general vaccine acceptance. As the popularity of social media as a whole has seemingly skyrocketed in the last decade, its use as a tool to disseminate information has had both beneficial and detrimental effects. In terms of efficiency, Reddit and other social media platforms have proven to be effective vehicles for information diffusion [3]. As a result, many studies have utilized the vast amount of data available on social media platforms and applied it to the field of sentiment analysis in order to gauge public opinion. While other studies have used sentiment analysis on Reddit data within the COVID-19 context [5,6,7,11], none have provided a state by state stratified analysis to assess the correlation between sentiment analysis metrics of COVID-19 vaccine related posts and COVID-19 vaccination rates in the United States. In this study, sentiment analysis is used on top-level Reddit comments from subreddits for each state in the United States. It is important to conduct this study on a more granular level (by state), since it has been shown that local knowledge and usage of social media by cities is directly related with a city's ability to respond to disaster [13]. Therefore, this study is necessary to completely understand the ways in which current natural language processing and sentiment analysis techniques can be used to gauge vaccine acceptance in the United States.

Author's address: Jake Cupani, jcupani1@umd.edu, University of Maryland, 7950 Baltimore Avenue, College Park, Maryland, USA, 20740.

0730-0301/2022/8-ART111 $15.00
https://doi.org/XXXXXXX.XXXXXXX

3 RELATED WORKS

3.1 Sentiment Analysis

Sentiment analysis is a field of data analysis and text mining that attempts to extract quantitative values, such as polarity (positive or negative) or subjectivity (opinion or fact), from text. With the advent of social media, sentiment analysis can now be used to understand the feelings, opinions, and behaviors of users on the specific topics discussed on these platforms. Classifying text as either positive, negative, or neutral can be boiled down to a classic classification problem where some model is used to predict this value as either continuous (any number -1 to 1) or multi-class (only -1, 0, or 1). Although polarity classification is a rather simple method of understanding the complexities of an opinion, it has seen success in previous works. Pang et. al used polarity classification on movie reviews as early as 2002 and similarly with work by Turney [20]. Many of the most popular of the social media networks, Twitter, Facebook, and Reddit, are frequently utilized in sentiment analysis studies due to the large volume of text data and accessibility to APIs

for data collection. In fact, in the survey of the sentiment analysis field done by Yue et. al, many benchmark datasets by Twitter, Facebook, and Amazon have been noted as being extensively used in research [20]. Given the sheer amount of data on these platforms, it is to no surprise that they could be leveraged in the field of sentiment analysis. However, as we will see, not all data on social media platforms is as easily accessible via APIs as others.

3.2 COVID-19 Sentiment Analysis on Social Media

Extension of sentiment analysis to COVID-19 related Reddit posts has actually been shown to be successful in multiple studies. Yan et al. found that the number of comments between July 13, 2020, and June 14, 2021 in the Vancouver subreddit were positively correlated with new COVID-19 cases in British Columbia [7]. The Canadian study conducts a similar stratified analysis, however the same is not done in the United States. Furthermore, Yan et al. uses a trained Random Forest model to classify comments into specific "topics" or general emotions such as "Joy", "Anger", "Sadness", and "Fear". While this is certainly an interesting approach, this study is fundamentally different in that it assesses the distribution of comment polarities, which provides a more aggregated approach. Other studies, such as Yin et al., utilize a sentiment analysis based method known as VADER to understand the attitudes of people towards the vaccine around the world [12]. Additionally, Low et al. showed how sentiment analysis and natural language processing could be used in analyzing the number of posts in health advice subreddits during the pandemic to identify loneliness and mental health related issues [5]. Understanding mental health conditions in the midst of the pandemic could play a role in overall vaccine acceptance and understanding. However, it is important to understand the reasons why these studies use Reddit compared to other social media platforms.

3.3 Benefits to using Reddit for Sentiment Analysis

Reddit in particular is unique from other social media in that subreddits can be used to automatically gather posts and comments relating to a specific topic. Other platforms, such as Twitter only have features such as lists and hashtags to sort tweets, making grouping of data collected rather intensive. The need for extra data preprocessing is done away with using Reddit data, since subreddits are inherently organized to contain only information pertaining to the designated topic. It also is relatively easy to gather information using the publicly available API. The ease of accessibility to social media data that Reddit provides is extremely useful to researchers. In particular, social media data can often be hard to access, since not all social media sites have an API or allow for their sites to be crawled [14]. These unique features are what makes using Reddit for sentiment analysis on specific topics so appealing.

4 RESEARCH QUESTION

This study aims to provide a stratified analysis to understand how sentiment analysis on Reddit top-level comments can be used as a proxy for COVID-19 vaccine acceptance. This study uses a combination of gathered Reddit data and CDC COVID-19 vaccination rates to understand the relationship between sentiment analysis metrics and vaccination rates. The main research question addressed in this study is:

(1) Can Reddit comment sentiment analysis be used as a proxy for COVID-19 vaccine acceptance on a state level?

By conducting a correlation analysis on sentiment scores and average vaccination rates by state, we can determine if applying sentiment analysis techniques to top-level Reddit comments is sufficient to gauge the public sentiment of vaccines by state. This study can help health organizations and administrations understand which states in the United States require more COVID-19 vaccination acceptance, education, and encouragement to boost vaccination rates.

5 DATASETS AND DATA COLLECTION

5.1 Reddit Datasets

The data collected from Reddit has been synthesized using PRAW, the Python Reddit API Wrapper. This Python package allows for easy data collection from the official Reddit API with features for extracting Reddit submission objects and comment forests for analysis. To extract the relevant COVID-19 posts from each state subreddit, all top posts were obtained from each state's subreddit. This dataset initially consisted of 33,768 unique comments across 1,621 unique posts. This dataset was then preprocessed and cleaned. This dataset was then filtered down to be within the timeframe of January 1st, 2020 to March 3rd, 2022, corresponding to the beginning of the COVID-19 pandemic to the time of this paper. The dataset was further filtered to contain the keywords "covid" or "vaccine" in the post title. The text in each comment in the dataset is cleaned by removing leading and trailing whitespace, special characters (,@,,<,>,",*,(,),/), and new line characters. Rows of comments that have the text "[deleted]" or "[removed]" also had to be dropped from the dataset, since these removed/deleted comments do not provide any analytical insight. After cleaning the data, there remains 22,824 unique comments from 1,615 unique posts. This dataset will be referred to as the "Full Dataset". Another dataset was created from this dataset for GPT-3 polarity scoring. Due to limited resources, only a subset of all of the comments could be used with the OpenAI API. As such, the new dataset is subsetted data between the dates of August 1st 2021 to March 1st 2022. This date corresponds to when approximately half of the total U.S. population was fully vaccinated, as seen in Figure 7. However, this dataset only contains 7,061 unique comments across 415 unique posts. This dataset will be referred to as the "August Dataset". Basic statistics for the larger and smaller dataset are seen in Figures 1 and 2.

5.2 COVID-19 Vaccinations Dataset

The datasets used for vaccination rates come directly from the Centers for Disease Control and Prevention's open data portal. This data is publicly available for anyone to access and was created on May 24th 2021 (available here). Administered vaccines are reported either by distributor or in aggregate allowing for greater granularity in analysis. In this work, vaccinations (fully vaccinated) as percent of population is used as the measurement of vaccination rate. State jurisdiction and date of reporting are also captured in this dataset. In order to prevent any double counting errors, aggregate

rows were filtered out by removing rows that did not contain a U.S. state abbreviation in the "Location" column. This effectively filters out the aggregate rows that are labeled as "U.S." in the Location column, as well as any rows pertaining to territories, which are not looked at in this analysis. The vaccination rates were then grouped by state for average vaccination rate by state. This is used in this analysis to provide a one-time snapshot of the state's overall vaccine acceptance.

6 METHODS

Three different techniques are used in order to discover which method of comment polarity scoring is most effective in this scenario. These techniques include basic TextBlob polarity scoring, title and comment polarity matching, and GPT-3 polarity scoring. Due to resource limitations, GPT-3 scoring was only done on comments in the August dataset.

6.1 TextBlob Polarity Scoring

TextBlob is a package for the Python programming language that offers an easy to use implementation of frequently used natural language processing techniques [18]. TextBlob can be used to identify parts of speech, extract noun phrases, and calculate polarity and subjectivity scores. This study focuses on utilizing the polarity and subjectivity scoring functions of the TextBlob library. First, comment text and post title text are iterated through and passed into the TextBlob function to create a TextBlob object for each row in the dataset. Next, the "sentiment" method of each TextBlob object is called to return a "Sentiment" object. This object can then be used to finally place the polarity and subjectivity scores into their respective columns. For this first technique, the data is grouped by state with the average polarity and average vaccination rates to provide a one-time snapshot of sentiment to vaccine acceptance. The same is done for title polarity. These values are then correlated using the .corr() function in Python to obtain the correlation coefficients. Correlations in this study use the standard Pearson correlation coefficient. Correlations are also done for both the Full and August datasets.

6.2 Title and Comment TextBlob Polarity Matching

The second technique used in this study is title and comment polarity matching. This technique attempts to compute a metric for how "antivax" a given comment is. For example, if a title that opposes the COVID-19 vaccine receives a negative title polarity, but a comment receives a positive polarity, then this would indicate that the user who commented agrees with the headline of the post. On the other hand, if a title has a positive polarity and a negative comment polarity, this would indicate that the user who commented does not agree with the title that is pro-vaccination. Given these discrepancies, each polarity score is thus inverted (multiplied by -1) if the title polarity is negative and comment polarity is negative, or if the title polarity is negative and the comment polarity is positive.

6.3 GPT-3 Polarity Scoring

The third and final technique used in this study is GPT-3 polarity scoring. This technique utilizes the GPT-3 model from the OpenAI API. GPT (Generative Pre-Trained Transformer)-3 is a deep learning neural network developed by OpenAi, one of the most prestigious artificial intelligence and machine learning research labs in the world. This model uses over 175 billion parameters and has been used in applications such as text-generation, Github Copilot, chatbots, and much more [21]. Since GPT-3 is able to be given a prompt in plain english, the GPT-3 model is passed a string containing the prompt "Decide whether a comment's sentiment is positive, neutral, or negative. Comment:", and the corresponding comment string. The GPT-3 model then outputs labels of Positive, Negative, or Neutral given the comment text. These labels are then converted to numeric values 1, -1, and 0, respectively. The correlation test between the average of these values by state and the average vaccination rates by state is then calculated for the correlation coefficient.

7 EXPERIMENTAL EVALUATION

Correlations tests were done to calculate correlation coefficients between the average polarity by state and average vaccination rate by state. These tests were done for each different technique and for each dataset, respectively. The results from these tests can be seen in Figure 6. The first correlation test conducted is between average comment polarity and average percent of the population fully vaccinated in the Full dataset. Fully vaccinated in this context refers to those that have a second dose of a two-dose vaccine or one dose of a single-dose vaccine. As seen in Figure 6, the correlation in the Full dataset between average comment polarity and average percentage of the population fully vaccinated using the basic TextBlob technique is around 0.5. On the other hand, the correlation coefficient between average title polarity and average vaccination rate by state using the same technique is lower around 0.31. Similarly, the same test done on the August dataset yields a correlation of 0.18 for titles and around 0.40 for comments. This indicates that the TextBlob polarity scoring techniques are better suited for comments than for titles. The same, however, cannot be said for the matching technique, which correlations ranged from 0.03 on the August dataset to 0.23 on the Full dataset. This is still interesting though as it shows that although the technique might not be the most effective, correlations across the board increased with sample size. Lastly, we have the correlation test done on the August dataset for the GPT-3 technique. This yielded a correlation coefficient of approximately 0.42, indicating that the GPT-3 model performs better than the TextBlob technique on the August dataset. Lastly, this work could be used by policymakers and health institutions to gauge what states in the United States are more or less hesitant to the vaccine and subsequently where to increase education, awareness, and promotion of vaccines to improve vaccine uptake.

8 LIMITATIONS AND FUTURE WORK

As with any research, there are limitations that need to be addressed. While this study is comprehensive in its analysis, there are plenty of ways in which it could be improved. For example, due to resource constraints, the GPT-3 scored comments used a max_tokens of 60, which means the GPT-3 model is able to tokenize around the first 45 words of the given comment. Setting a higher value for max_tokens would increase the overall cost of scoring all comments in the dataset

since the OpenAI API is a pay-as-you-go model. Additionally, it is important to keep in mind that Reddit users are predominantly younger, white, and male [18], which does not accurately represent the broader demographics of the United States. Other papers have also pointed out the limitation of the TextBlob package, which might perform worse than highly tuned models specific to COVID-19 discussion [7]. However, given that the GPT-3 model is extremely robust and trained on an extensive amount of data, it should be able to classify comments with accuracy comparable, if not better, than other simpler, fine tuned models. Future work on this specific study could utilize fine tuning to COVID-19 discussion by using OpenAI's fine tuning features in the OpenAI API. These features allow for higher quality results by providing training data directly to the API [19]. Future work could also investigate this further on the Full dataset. This study could also be expanded into a time series analysis in which sentiment scores are calculated for comments during specific time periods and correlated by time frame. This would provide a deeper, more granular look into COVID-19 vaccine sentiment. More data could also be gathered from the Reddit API to increase the sample size, since one limitation of this study is that some states had more comments than others.

9 CONCLUSION

By exploring different ways of scoring Reddit top-level comment polarities across the United States, this study has shown that utilizing the TextBlob package and state of the art models such as GPT-3 retains a high correlation between comment polarities and average vaccination rates. While the ad-hoc method of comparing comment polarities to title polarities did not perform well, both the TextBlob and GPT-3 scoring techniques showed similar promising results. These results indicate that polarity based sentiment analysis of Reddit top-level comments can be used as a proxy for vaccine acceptance on the state level.

10 FIGURES

	title_polarity	title_subjectivity	comment_polarity	comment_subjectivity
count	22824.000000	22824.000000	22824.000000	22824.000000
mean	0.039794	0.266788	0.072863	0.545841
std	0.223588	0.278787	0.333330	0.211885
min	-1.000000	0.000000	-1.000000	0.000000
25%	0.000000	0.000000	-0.113659	0.400000
50%	0.000000	0.227273	0.087500	0.531250
75%	0.136364	0.483333	0.250000	0.666667
max	1.000000	1.000000	1.000000	1.000000

Fig. 1. Reddit Full Dataset Statistics

11 REFERENCES

(1) Safety of covid-19 vaccines. (n.d.). Retrieved March 17, 2022, from www.cdc.gov/coronavirus/2019-ncov/vaccines/safety/safety-of-vaccines.html

(2) Covid-19 vaccine confidence. (2022, February 28). Retrieved March 17, 2022, from https://www.cdc.gov/vaccines/covid-19/vaccinate-with-confidence.html

	post_num_comments	title_polarity	comment_polarity	title_subjectivity	comment_subjectivity	sentiment_scores
count	7061.000000	7061.000000	7061.000000	7061.000000	7061.000000	7061.000000
mean	218.654298	0.021393	0.060127	0.251762	0.556341	-0.652740
std	122.495277	0.216334	0.350950	0.267763	0.211254	0.644493
min	1.000000	-1.000000	-1.000000	0.000000	0.000000	-1.000000
25%	117.000000	0.000000	-0.146605	0.000000	0.412500	-1.000000
50%	210.000000	0.000000	0.072222	0.166667	0.541667	-1.000000
75%	301.000000	0.118182	0.250000	0.497273	0.678571	0.000000
max	589.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Fig. 2. GPT-3 Dataset Statistics

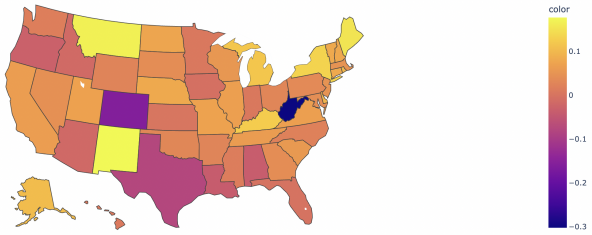

Fig. 3. Average Title Polarity by State

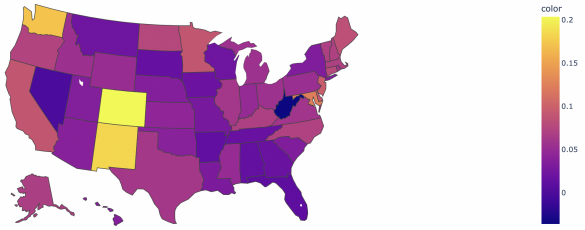

Fig. 4. Average Comment Polarity by State

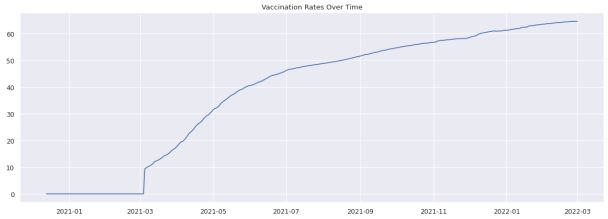

Fig. 5. Percent Fully Vaccinated Over Time

	TextBlob	GPT-3
Full Dataset (Title)	0.310446	NA
Full Dataset (Comments)	0.500573	NA
Full Dataset (Matching)	0.231466	NA
August Dataset (Title)	0.18069	NA
August Dataset (Comments)	0.397819	0.417077
August Dataset (Matching)	0.038456	NA

Fig. 6. Average Comment and Title Polarity to Average Vaccination Rate Correlations by Dataset

(3) Guille, A., Hacid, H., Favre, C., Zighed, D. A. (2013). Information diffusion in online social networks. ACM SIGMOD Record,
42(2), 17-28. doi:10.1145/2503792.2503797
(4) Loomba, S., De Figueiredo, A., Piatek, S. J., De Graaf, K., & Larson, H. J. (2021). Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. Nature Human Behaviour, 5(3), 337-348. doi:10.1038/s41562-021-01056-1
(5) Low, D. M., Rumker, L., Talkar, T., Torous, J., Cecchi, G., & Ghosh, S. S. (2020). Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during COVID-19: Observational study. Journal of Medical Internet Research, 22(10). doi:10.2196/22635
(6) Melton, C. A., Olusanya, O. A., Ammar, N., Shaban-Nejad, A. (2021). Public sentiment analysis and topic modeling regarding COVID-19 vaccines on the Reddit social media platform: A call to action for strengthening vaccine confidence. Journal of Infection and Public Health, 14(10), 1505-1512. doi:10.1016/j.jiph.2021.08.010
(7) Yan, C., Law, M., Nguyen, S., Cheung, J., & Kong, J. (2021). Comparing public sentiment toward covid-19 vaccines across Canadian cities: Analysis of comments on Reddit. Journal of Medical Internet Research, 23(9). doi:10.2196/32685
(8) COVID-19 Vaccinations in the United States. (2022, March 17). Retrieved March 17, 2022, from https://covid.cdc.gov/covid-data-tracker/#vaccinations_vacc-people-onedose-pop-5yr
(9) Report of the sage - World Health Organization. (2014, October 1). Retrieved March 18, 2022
(10) Alamoodi, A. H., Zaidan, B. B., Zaidan, A. A., Albahri, O. S., Mohammed, K. I., Malik, R. Q., ... Alaa, M. (2021). Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review. Expert systems with applications, 167, 114155.
(11) Murray, C., Mitchell, L., Tuke, J., Mackay, M. (2020). Symptom extraction from the narratives of personal experiences with COVID-19 on Reddit. arXiv preprint arXiv:2005.10454.
(12) Yin, H., Song, X., Yang, S., Li, J. (2022). Sentiment analysis and topic modeling for COVID-19 vaccine discussions. World Wide Web, 1-17.
(13) Kim, J., Hastak, M. (2018). Social network analysis: Characteristics of online social networks after a disaster. International journal of information management, 38(1), 86-96.
(14) Alamoodi, A. H., Zaidan, B. B., Al-Masawa, M., Taresh, S. M., Noman, S., Ahmaro, I. Y., ... Salahaldin, A. (2021). Multi-perspectives systematic review on the applications of sentiment analysis for vaccine hesitancy. Computers in Biology and Medicine, 139, 104957.
(15) Covid Data Tracker. (2022, April 14). Retrieved April 14, 2022, from https://covid.cdc.gov/covid-data-tracker/#datatracker-home
(16) COVID-19 Vaccinations in the United States,Jurisdiction. (2022, April 14), from https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-Jurisdi/unsk-b7fc
(17) TextBlob: Simplified Text Processing. (2022, April 14), from https://textblob.readthedocs.io/en/dev/index.html
(18) Chipidza W. (2021). The effect of toxicity on COVID-19 news network formation in political subcommunities on Reddit: An affiliation network approach. International journal of information management, 61, 102397. https://doi.org/10.1016/j.ijinfomgt.2021.102397
(19) Open AI Fine Tuning. (2022, May 10). Retrieved May 10, 2022, from https://beta.openai.com/docs/guides/fine-tuning
(20) Yue, L., Chen, W., Li, X., Zuo, W., Yin, M. (2019). A survey of sentiment analysis in social media. Knowledge and Information Systems, 60(2), 617-663.
(21) What is GPT-3? (2021, June), from www.techtarget.com/searchenterpriseai/definition/GPT-3