Pfizer Vaccine Tweets: A Sentiment Analysis

Justin Ehly, Nicole Norelli, and Mingyang Nick YU

Abstract—A sentiment analysis of tweets selected using the #PfizerBioNTech hashtag posted between December 12, 2020 and March 13, 2021 was conducted. Location information provided by Twitter users was identified and utilized. Sentiment analysis was obtained through Amazon Comprehend natural language processing, and an examination of sentiment over time showed an increasing proportion of positive sentiment. Possible small regional differences in sentiment within the United States were detected; however, no relationship between sentiment and COVID-19 death rates or other twitter user information was found.

Index Terms—Data and knowledge visualization, Data Communications, Direct data manipulation, Sentiment analysis

1 Introduction

WORLDWIDE distribution of the Pfizer vaccine in response to the COVID-19 pandemic has increased, but public perceptions about the vaccine and vaccine hesitancy remain as potential barriers to achieving herd immunity in many countries [1]. Trust in healthcare entities and government are contributing factors to a population's confidence in vaccines [2], and misinformation about vaccines, particularly from online sources, can negatively impact vaccine intent [1], [3].

Analysis of public sentiment could assist the medical and policy communities in developing a more effective informational campaign, which would save lives. Determining sentiment trends related to time, geography, or particular events would be helpful in planning such campaigns. This project analyzed real-world Twitter data related to the Pfizer vaccine to identify public perceptions of the vaccine as well as any useful trends that might inform public policy.

2 Data Description

2.1 Data Set

The data set for our analysis was supplied by Kaggle. It consisted of 16 variables from 6,821 tweets which were collected using the tweepy Python package and Twitter API. The tweets were selected using the #PfizerBioNTech hashtag, and they were posted between December 12, 2020 and March 13, 2021. We analyzed the Twitter data in two sections. Our initial data set consisted of 4,139 tweets posted between December 12, 2020 and January 20, 2021. Our full data set consisted of the initial set plus the additional 2,682 tweets posted between January 20, 2010 and March 13, 2021.

2.2 Missing Data

The variables with missing data from our full data set were: user_description (427), user_location (1381), and hashtags (1676). Additionally, the user_location data varied in format, with users identifying country, state, city and state, city and country, city, region, or fictional locations. The format of states also included abbreviations as well as full state names.

2.3 Geographic Variables

Because we were interested in geographic trends, we prioritized transforming location data into a consistent format. Using our initial data set, we created variables for user_country, user_city, user_state, and us_state_code. We also obtained data sets with lists of world countries, world cities, and U.S. states.

First we wrote a function to match the world city names iteratively, descending by string length. Once we matched a city, the user_country variable was filled in according to the corresponding country from the world city data set. After we completed the process, we identified all short and unfamiliar city names with high frequencies. Many were erroneous matches, and we treated them as special cases, which we resolved individually. For example, city "ARTH" was often matched to "EARTH" and "NADA" was often matched to "CANADA."

Next, we used the world country names data set to identify and fill in the user_country variable where still necessary. Then, we matched U.S. states, checking to make sure the country name was correct when a state was matched. We identified particular formats and cases that resulted in errors, such as nonstandard abbreviations (like NYC) or WEST VIRGINIA matching to VIRGINIA and fixed them accordingly. Additionally, we addressed location entries formatted in a "city, state code" structure to capture the state information. Finally, we supplied the appropriate state codes for the user_state_code variable.

While we applied this entire process to our initial data set, we decided to only identify user_state and us_state_code for the new tweets (January 20, 2010 to March 13, 2021) in the full data set, as state identification was our primary interest.

3 Sentiment Analysis

To analyze tweet sentiment, we first cleaned and refor-

- *J. Ehly is in the Master of Science in Data Science program, Southern Methodist University, Dallas, TX 75275. E-mail: jehly@ mail.smu.edu.*
- *N. Norelli is in the Master of Science in Data Science program, Southern Methodist University, Dallas, TX 75275. E-mail: nnorelli@ mail.smu.edu.*
- *M. YU is in the Master of Science in Data Science program, Southern Methodist University, Dallas, TX 75275. E-mail: nyu@ mail.smu.edu.*

matted the text variable. We then uploaded the tweet text to Amazon S3, an object storage service. Next, we utilized Amazon Comprehend to obtain a sentiment analysis. Amazon Comprehend uses natural language processing to determine the emotional sentiment in text. It provides a likelihood score for positive, negative, neutral, and mixed sentiment and identifies the most likely candidate for each document. (In this case, each tweet was considered a document.) The sum of the four scores equals to one, so each score can also be thought of as a probability. We then matched the sentiment information back to the original entry for our exploratory analysis.

4 Data Analysis

4.1 Geographic Trends: Country and City

We explored tweet sentiment by geographical location for user countries, states, and cities. There was a large amount of missing data, and there were many locations with only a handful of tweets, so each of the following sections only examines locations with ten or more tweets.

The city and country data are derived entirely from the initial data set, as we chose not to identify city or country when we added the most recent tweets to our full data set. Countries with no positive sentiment included: Cyprus, Denmark, Japan, Niger, Philippines, Qatar, Turkey, and Venezuela. These countries had between 11 and 43 tweets each, so the sample size was relatively small.

Fig. 1 shows the proportion of tweet sentiment by country for countries with 50 or more tweets. Ireland had the highest proportion of positive tweets, while India had the lowest. UAE had the smallest proportion of negative tweets. Overall, there was no apparent larger pattern regarding sentiment and country. Because we filled in user_country when we identified a city, there was substantial overlap between the city and country analyses. City sentiment trends followed the country trends previously identified.

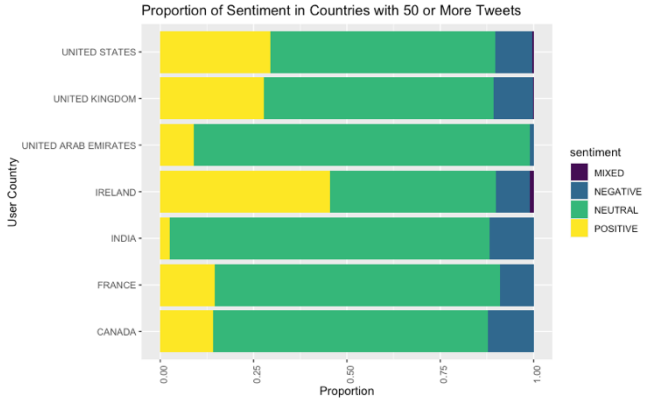

Fig. 1. Proportion of positive, neutral, negative, and mixed sentiment in countries with 50 or more tweets.

4.2 Geographic Trends: State

We identified state data for the full data set, as this was our primary area of interest when examining geographical trends. There were 953 tweets identified by state. In states with ten or more tweets, we found the highest proportion of positive tweets came from Colorado, and Vermont had no positive tweets at all. Interestingly, the Colorado tweets were from a variety of users, but several of them identified as medical professionals. Most of the Vermont tweets were from a single user, so little meaning can be attributed to its lack of positivity.

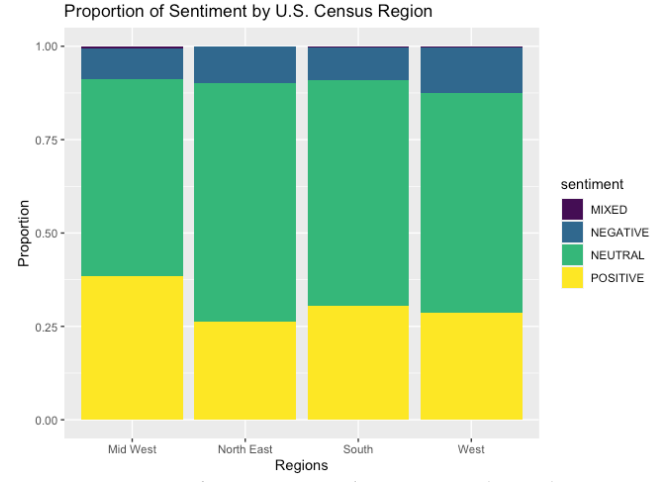

Fig. 2. Proportion of positive, neutral, negative, and mixed sentiment in U.S. census regions

4.3 Geographic Trends: Census Region

Because the average number of tweets in each state was relatively small, we also examined sentiment grouped by the four U.S. census regions. There appeared to be a slightly higher proportion of positive sentiment in the Midwest and a slightly higher proportion of negative sentiment in the West (Fig. 2).

4.4 Geographic Trends: Political Party

Due to the politicized nature of responses to COVID-19 in the United States, we wanted to examine sentiment as it related to political party. We chose to label each state as Republican or Democrat based upon its 2020 presidential election results. Then, we compared proportion of sentiment between these two groups. Surprisingly, sentiment proportions looked very similar between the two groups. Fig. 3 shows the similar proportions, with slightly higher negative sentiment in the Democratic group.

Of course, this method of assigning political party alignment was not ideal, as the individuals tweeting from each state may or may not identify with the majority politically.

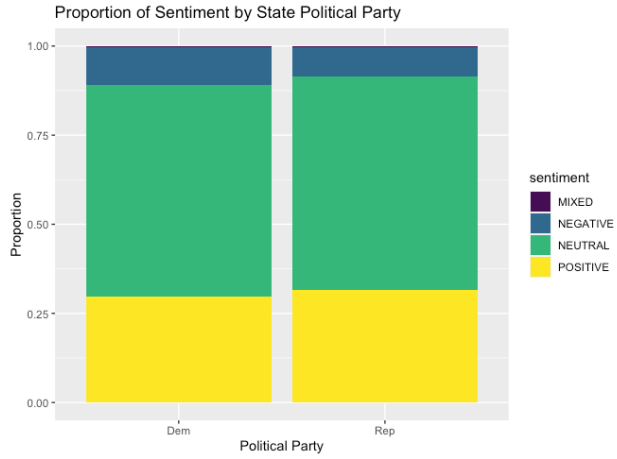

Fig. 3. Proportion of positive, neutral, negative, and mixed sentiment by state political party outcome in the 2020 presidential election

4.5 COVID-19 Cases by State

We compared sentiment analysis by state with COVID-19 case information by state. We obtained daily case information by state from the CDC and incorporated state population information from the U.S. Census to find the mean total case count in each state over the time period of the tweets.

There did not appear to be any meaningful overlap between COVID-19 case rates and sentiments within states. We identified the states with the highest proportions of negative and positive sentiment and looked for overlap with high or low total state case count. For example, the ten states with the lowest total case count contained both higher negative sentiment (VT, WA) and higher positive sentiment (MI) states. The ten states with the highest total case count only contained one higher proportion positive (WI) state. No clear relationships between total COVID-19 case numbers and sentiments were apparent.

4.6 Sentiment over Time

Previous research into vaccine hesitancy and the effects of social media shows that social media exposure tends to have an echo chamber effect. Users are primarily exposed to views similar to their own, and this leads to the polarization of views over time [4]. Our initial data set (December 12, 2020 to January 20, 2021) appeared to support this model. Fig. 4 shows the proportion of sentiment over time, with increasing proportions of positive and negative sentiment and corresponding decreasing proportions of neutral sentiment.

However, when we added the more recent tweets to our full data set analysis, we discovered a different trend. When we included data from January 20, 2021 to March 13, 2021, we found an overall increase in positive sentiment without the increase in negative sentiment (Fig. 5).

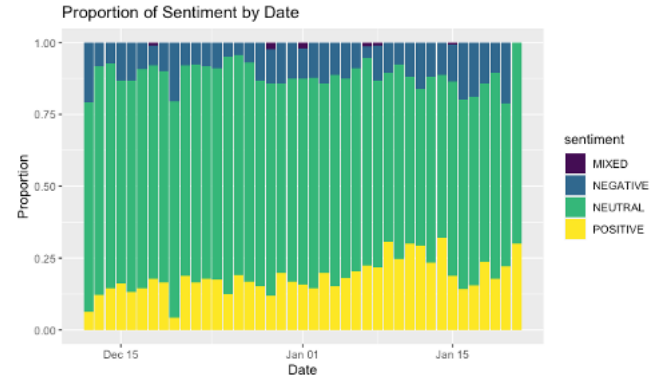

Fig. 4. Proportion of sentiment over time from December 12, 2020 to January 20, 2021 (initial data set).

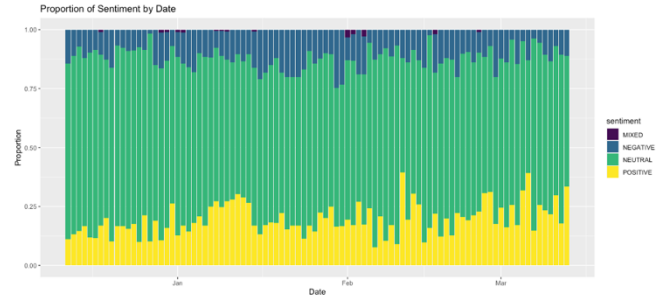

Fig. 5. Proportion of sentiment over time from December 12, 2020 to March 13, 2021 (full data set).

We also determined the mean likelihood of each sentiment by day. Fig. 6 illustrates that the mean likelihood of positive sentiment seems to be increasing whereas the mean likelihood of negative and neutral sentiment seem to be decreasing. This may indicate that sentiments are more clearly positive over time.

This increase was a surprising development, given the polarization effects that vaccine hesitancy research would predict. Increasing positive sentiment may be due to increasingly clear information about how the vaccine works, expected side effects, and clearer understanding of how to get the vaccine. Also, as more of the population has a close relationship with someone who has been vaccinated [5], negative sentiment might decrease.

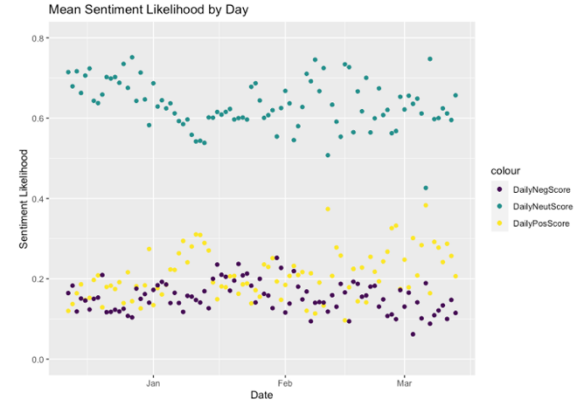

Fig. 6. Mean positive, negative, and neutral sentiment each day from December 12, 2020 to March 13, 2021 (full data set).

4.7 Other Twitter Account Information

We examined the relationship between sentiment and other Twitter user information, such as user_followers, retweets, user_friends, user_favourites, and date user_created. We failed to find any notable relationships between these user data variables and overall sentiment regarding the vaccine.

4.8 Hashtag Comparison: Positive vs. Negative

We wanted to identify words or phrases that were utilized by users under different sentiment groups. These words or phrases could potentially help narrow down reasons why certain users tend to have positive or negative sentiment. To accomplish this, we generated word clouds from hashtags that were captured with the hashtags variable using word frequencies under different sentiment groups.

The word cloud of hashtags from the positive sentiment group (Fig. 7) showed a high frequency of types of vaccine names and words such as: science, scientists, grateful, and firstdose.

Fig. 7. Word cloud of hashtags from positive sentiment tweets.

While the negative sentiment group (Fig. 8) also included a high frequency of hashtags regarding vaccine names, we also see words such as obesity, fitness, and location names. This leads us to hypothesize that users

Fig. 8. Word cloud of hashtags from negative sentiment tweets

with positive sentiment tend to express their appreciation for science, while users may express negative sentiment due to frustration with various health conditions, local vaccine availability, or towards a particular location rather than the vaccine itself.

This may set a direction for researchers to further study the root cause of positive or negative sentiment toward COVID-19 vaccines and help both domestic and foreign governments to overcome vaccination objections from the general public.

5 Conclusions

Overall results from our sentiment analysis found increasing positive sentiment over time, with a higher proportion of positive sentiment in the Midwest and a slightly higher proportion of negative sentiment in the West. Although the difference between census regions was relatively small, this could indicate that campaigns regarding the Pfizer vaccine should be concentrated more in the western areas of the United States for maximum effectiveness.

Additionally, the overall positive increase might indicate that the information available to the public regarding the Pfizer vaccine is having a positive effect. While this is just observational data, and no causal conclusions can be drawn, the improvement in sentiment over time is encouraging news for government and healthcare entities trying to improve public perceptions of the vaccine.

An examination of overall tweets per day (Fig. 9) shows the volume of tweets decreasing near the end of January or the beginning of February. This could be a natural decline as public attention waned, or it could be a response to changes in U.S. national policy and/or Biden administration messaging. Additionally, Johnson & Johnson announced its application for Emergency Use Authorization for its vaccine on February 4 and earned approval on February 28, which raised attention to alternate vaccines and could have reduced the volume of Pfizer-related tweets.

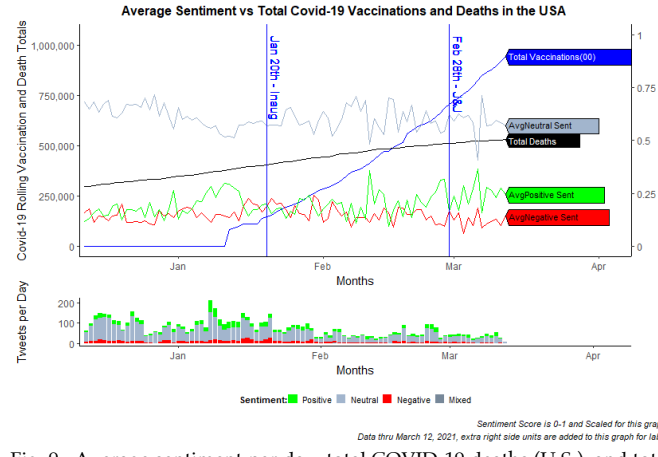

Fig. 9. Average sentiment per day, total COVID-19 deaths (U.S.), and total vaccinations (U.S.), with significant dates marked. Bottom: Number of tweets per day broken down by sentiment.

6 Limitations and Future Directions

6.1 Limitations

Of course, all user data attached to twitter accounts is self-reported. Even when we were able to extract the city, state, or country of a user, there is still the possibility that they did not report their location truthfully.

Because of the many users with unreported locations, our sample sizes for U.S. states were relatively small. Caution should be applied to any conclusions regarding trends in states.

This data set only included tweets with the #PfizerBioNTech hashtag. Now that a variety of COVID-19 vaccines have gained approval world-wide, expanding the hashtag requirement to include new vaccine names or more general vaccine terms might better illustrate general sentiment toward COVID-19 vaccines.

Our analysis only included English language tweets. This restricts the population from which we are sampling to be English-speaking twitter users who have used the #PfizerBioNTech hashtag in their tweets.

6.1 Future Directions

Expanding the analysis to a real-time framework would give more timely, up-to-date information about sentiment trends.

Also, expanding the hashtags used to pull tweets might result in a larger sample size. This would be useful when analyzing geographic trends as well as examining the potential relationship between sentiment and particular events.

An expansion to non-English tweets could also prove to be interesting. It is possible that sentiment trends could be different when other languages are analyzed.

7 Supplemental Information

For original data set, please see:
https://github.com/justinehly/7330-Term-Project/blob/main/Data%20Set/vaccination_tweets0313.csv
https://github.com/justinehly/7330-Term-Project/blob/main/Data%20Set/tweet_posts.csv
For data cleaning code, please see:
https://github.com/justinehly/7330-Term-Project/blob/main/PfizerVaccineAnalysis.RMD
https://github.com/justinehly/7330-Term-Project/blob/main/PfizerVaccineClean0313.RMD
For additional EDA, visualizations, and data analysis, please see:
https://github.com/justinehly/7330-Term-Project/blob/main/EDA_by_Nicole
https://github.com/justinehly/7330-Term-Project/blob/main/EDA_by_Nick.RMD
https://github.com/justinehly/7330-Term-Project/blob/main/Vaccinations_Ehly.Rmd

References

[1] Loomba, S., de Figueiredo, A., Piatek, S.J. et al. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. Nat Hum Behav (2021). https://doi.org/10.1038/s41562-021-01056-1W

[2] de Figueiredo, A., Simas, C., Karafillakis, E., Paterson, P. & Larson, H. J. Mapping global trends in vaccine confidence and investigating barriers to vaccine uptake: a large-scale retrospective temporal modelling study. Lancet 396, 898–908 (2020).

[3] Kim, H. K., Ahn, J., Atkinson, L. & Kahlor, L. A. Effects of COVID-19misinformation on information seeking, avoidance, and processing: a multicountry comparative study. Sci Commun 42, https://doi.org/10.1177/1075547020959670 (2020).

[4] Schmidt, A. L., Zollo, F., Scala, A., Betsch, C. & Quattrociocchi, W. Polarization of the vaccination debate on Facebook. *Vaccine* **36**, 3606–3612 (2018).

[5] KFF Health Tracking Poll/ KFF COVID-19 Vaccine Monitor (February 15-23, 2021)