Issues in Information Systems
Volume 24, Issue 4, pp. 115-129, 2023

DOI: https://doi.org/10.48009/4_iis_2023_110

Twitter analysis on COVID-19 vaccine sentiment in February of 2021

David Curtis, *University of Charleston, West Virginia, davidcurtis@ucwv.edu*
Jehu Humphries, *University of Charleston, West Virginia, jehuhumphries@ucwv.edu*
Taiwo Ajani, *University of Charleston, West Virginia, taiwoajani@ucwv.edu*

Abstract

This paper provides an investigative summary of U.S. Twitter user sentiment on the availability of COVID-19 vaccines during the period of February 03-10, 2021. The sentiments were captured from 2,000 Twitter Tweet data observations collected during a seven-day period. This period of time was when the COVID-19 pandemic had reached one-year maturity, and the spread of the virus was showing a gradual decline of daily cases since peaking at over 300,000 on January 8, 2021 (CDC, 2021). Additionally, during this time, a pandemic milestone had been reached in which 34 million vaccinations (10% of the U.S. population) had been administered. The intent of this paper is to depict the tone of conversation about the COVID-19 vaccine through sentiment analysis, and to determine if sentiment scores were highly impacted by the location that the Tweet was authored. This is an exploratory study to provide baseline information in a subject area that is still in its infancy relative to other subject areas in the sentiment analysis discipline. The 2,000 Twitter Tweet data observations were captured from a sample representing every state in the U.S. and provides a broad-spectrum representation of sentiment from people of many diverse geopolitical, socioeconomic, scientific, medical backgrounds.

Keywords: tweet analysis, vaccine, pandemic, COVID-19, immunity, sentiment, distribution

Introduction

On December 12, 2019 the COVID-19 virus entered the global arena from China's Hubei Province (CDC2, n.d.). On January 20, 2020, the first confirmed cases of infection in the U.S. came from patient samples recovered in the State of Washington. From that point, the disease spread throughout the U.S. and became the global pandemic as we know it. About eleven months after it was first discovered in the US, the first vaccine was unveiled (December 11, 2020, under the Federal Food and Drug Administration (FDA) Emergency Use Authorization (EUA) act (CDC2, n.d). Subsequently, the U.S. sentiment on drug efficacy, availability, and production method fluctuated drastically within the broad spectrum of very positive (Twitter Sentiment Scale: 1) to very negative (Twitter Sentiment Scale: -1). On February 11, 2021 the U.S. Government achieved a 10% population vaccination milestone after peaking at over 300,000 cases per day a month earlier. The Twitter data gathered for this study was collected just one week prior to the U.S. Government achieving the 10% population vaccination milestone. What was the U.S. population sentiment, for and against the Corona Virus vaccine during this period of time?

The purpose of this study is to measure and describe the sentiment of the U.S. population on the vaccine with analytical modeling utilizing a broad-spectrum sample of Twitter Data gathered from February 3rd through 10th of 2021. Additionally, the process of data cleaning or 'munging' is described in detail in this paper to provide the reader an example of how to extract valuable information from datasets not yet ready for data analysis.

Literature Review

Although the entrance of the Corona Virus onto the world scene may seem enigmatic or difficult to understand, the impact of the virus on the entire global population has been dramatic and very traumatic. As of the writing of this paper, the U.S. experienced over 1.6 million deaths attributed to COVID-19, and this is 24% of the total 6.55 million people lost globally as of October 6, 2022 (The Washington Post, 2022). While the death toll experienced throughout the world is a stark variable which will be viewed in the future as a primary statistic for crisis assessments, the pain and suffering inflicted on survivors and caretakers cannot be ignored. Contracting the virus or supporting someone who has contracted the virus while trying to maintain a certain amount of separation from the virus can be a traumatic experience.

The traumatic experience of survivors including reported near-death experiences could have long-lasting or lifelong repercussions. The consequences of catching the virus and dying or experiencing an anguishing period of sickness and recovery are not the only consequences or long-term dilemmas the Coronavirus epidemic created. The socioeconomic and geopolitical impacts from the Coronavirus also have consequences on relationships and social interactions within and outside the United States (Samuel et al., 2020a). Whether it is intermingling restrictions, flight cancellations, or referendums to prevent further infections from reaching the states or other countries, the socioeconomic and geopolitical impacts have been widespread and rampant.

Therefore, it is not a surprise that these socio-public sentiments are expressed on several platforms in our social media driven world. Such sentiments portrayed on social media platforms such as Twitter and Facebook, have fluctuated drastically to reflect the severity and brutal nature of the impact of the pandemic on world society. Studies such as this one, have been crafted using various analytical platforms to measure public feeling or sentiment on a wide range of societal issues with the premise of applying analytics to public sentiment scenarios (PSS). The goal is to understand potential reactions to public issues which may have a socioeconomic or geopolitical effect on governmental policy, the economy, or even acceptance or rejection of the COVID-19 vaccine rollout program (Samuel et al., 2020a). Traumatic and tragic events such as COVID, will periodically occur within the human timeline of experience, and the impact of these scenarios cannot be wished away. Understanding sentiment fluctuations both positive and negative within PSS networks can inform governmental as well as business conglomerates of potential avenues to minimize the potential impact of traumatic and tragic events. Examining these trends can also potentially aide care givers to provide assistance more quickly based on understanding what is hurting or helping people. Measuring sentiment sway within PSS networks is a powerful method to understand what people want and need in times of crisis.

Similar to previous studies on the COVID-19 pandemic conducted by Rahman et al. (2021), this study aims to find factors that contribute to a sentiment score associated with a tweet. In previous studies, sentiment is viewed as the sole factor to determine public opinion with a narrowed focus to determine if socioeconomic factors are a driver for sentiment. This study will use similar tools and processes to generate conclusions without referencing socioeconomic information. The only available information for analysis in this study are the variables contained within the data set. Rahman and his coauthors (2021) describe common cleaning techniques required for R analysis of twitter data. The dataset utilized in this study was cleaned similarly to remove stop words, tokenize the words within the text, and then provide a sentiment score where the data could be grouped, filtered, and analyzed.

Sentiment analysis or opinion mining is an analytical method designed to apply natural language processing (NLP) to determine and scale positive, neutral or negative intonation in text (Barney, n.d., Pang & Lee, 2008). The practice of sentiment analysis is used widely in many industries to gauge responses or opinions of consumers and customers, and is utilized to inform future organizational business decisions. Textual analysis can also be performed to measure the intent behind a contributor message to determine whether they are asking a question, making a complaint, making a suggestion, etc… (Gupta, 2018). Many different types and levels of textual analysis can be algorithmically applied to derive insights from peoples' messages. Textual analysis can be performed on single words (unigrams) or emoticons, and even multiple word combinations (Agarwal, 2011).

Measuring word sentiment is based upon a scaled lexicon of words which have a positive to negative value applied to them. The programming language used in this study was Posit R which contains several lexicons or dictionaries which can be selected based on the type of analysis conducted. Some of the lexicons utilized in R for textual analysis are the AFINN, Bing, Jockers, NRC, Harvard-IV, Henry's Financial and Loughran-McDonald dictionaries (Hill, n.d., Feuerriegel & Proellochs, 2021). Each lexicon differs in its output based on the values associated to the words. Some lexicons provide scoring based integer values, some grade each word as positive, neutral or negative, and others equate the words to emotional descriptors. Lexicon selection is based on the type of analysis to be performed. As an example, the AFINN lexicon or dictionary provides an integer value between 5 and -5. A grade of 5 represents the highest value of positive, 0 represents neutral, and -5 represents the most negative word sentiment (Silge & Robinson, n.d). Selection of the right lexicon can greatly enhance a researcher's ability to perform both simple and complex analytical text calculations.

Methodology

This study provides exploratory analysis on the Twitter dataset and tests the hypothesis that the mean sentiment score is equal in all 50 US States. The null hypothesis is that the mean sentiment score is equal in all 50 States, and the alternative hypothesis is that the mean sentiment score differs in at least one State. For the purposes of the study, the research team will temporarily assume equal sentiment values until the assumption can be statistically tested.

The project team conducted primary aspects of sentiment analysis similar to Samuel et.al. (2020b). Societal sentiment as projected by the American populace within social media platforms such as Twitter is measured based on key words reflecting positive, negative or neutral intonation (Samuel et al., 2020b). Textual visual analysis is also performed to reflect 250 key words sorted by primacy in the form of *Word Clouds* which portray overall levels of sentiment values intonating positive, negative, or neutral sentiment. The techniques to perform the textual analysis used in this study were in line with natural language processing (NLP) techniques but were not exhaustive in form to compare to other methods used in similar studies (Samuel et al., 2022b). Included in this study are Linear Modeling techniques to correlate dependent variable relationships with independent variables, and analysis variance (ANOVA) testing of validity. In line with the referenced studies mentioned above, the development team worked diligently to implement and utilize analytical methods consistent with respected studies similar in form and context. The project team is confident the material and conclusions contained within will benefit prosperity and the pursuit of truth in analytical exploratory and descriptive testing.

The research team followed a deliberate process to manipulate the raw data pulled from Twitter to conduct exploratory analysis and hypothesis testing. The first step in the process was to clean the raw data only to include relevant information in the study. We formatted the remaining variables to ensure accurate analysis. We conducted exploratory processes to group the data based on their values to analyze the dataset further. With a clean data set, the research team next conducted textual analysis to understand better what language created a sentiment score and the words most frequently associated with each type of observation. Finally, we tested the hypothesis that all states maintain the same sentiment or tone when exchanging ideas and opinions about the COVID-19 vaccine.

To execute sentiment analysis, all calculations and analysis were conducted in the R programming language. The research team leveraged the "sentiment analysis" package to tokenize or separate each word into individual linguistic units. Next, the research team removed the common connectors for nouns, verbs and adjectives (stop words), and the symbols and graphics from each observation to ensure analysis was conducted only on words contributing to the main idea and emotion contained within the language used in the tweet. Once each word was tokenized, a sentiment score was applied to each word through a function in R. The function cross referenced the words remaining in the observation against a dictionary containing a list of positive and negative words according to the psychological Harvard-IV dictionary as used in the General Inquirer (GI) software (Proellochs, 2021). While the sentiment analysis package contains the ability to reference other dictionaries, such as Henry's financial dictionary and the Loughran-McDonald Financial dictionary, it was the assessment that the score contained within Harvard-IV dictionary would be the most applicable to the study. For that reason, all sentiment scores in this study represent values created by the sentiment analysis package referencing the sentiment GI variable within the data set.

The result of the process allowed the research team to then bundle each observation into categories defined by a numeric value describing the tone of speech in the observation as either positive, negative, or neutral. From the scoring of each observation, further exploratory analysis allowed the research team to package each observation contained within the geographic boundaries of a U.S. State to create an average sentiment score for each territory. The sentiment score assigned to each State accounts for each observation throughout the collection period to capture fluctuations in the frequency of tweets in a region and collect shifting tones as varying opinions and ideas are expressed in a public forum. The mean sentiment score and the totality of observations from each state are the benchmark values for hypothesis testing through the ANOVA test.

Results

The five focus areas for the research team are the sentiment variable, the type of tweet variable, the date variable, the State variable, and the text variable. By focusing exploratory efforts on the five variables pertinent to the study, the research team worked towards forming conclusions from the dataset. During the exploratory process, the research team first developed an understanding of the sentiment variable and then explored the date variable independent of one another. Next, the study required grouping the sentiment for every observation with its paired US State. Following the grouped analysis of sentiment and State, the study then explored the grouped sentiment and date variables throughout the data set. After completion, the team sought to understand the representation of each type of observation; positive, negative, and neutral. The last step of the exploratory phase of the dataset concludes with a textual analysis of each type of tweet. Finally, the research team finished the dataset analysis by conducting an ANOVA test to challenge the hypothesis that the mean sentiment value was equal in all US States.

Issues in Information Systems
Volume 24, Issue 4, pp. 115-129, 2023

Sentiment Analysis

A boxplot was used to explore the values within the variables. It was observed that the mean sentiment for the COVID-19 vaccine was 0.05379, and a significant number of outliers were located away from the mean (Figure 1). The output of the boxplot before data cleaning efforts represents all 2,000 observations in the data set. The boxplot output after cleaning is also shown below (Figure 2). The mean sentiment value decreased from 0.05379 to 0.03579. We concluded that although small, the overall positive sentiment from advertisements and bots was significant enough to shift the mean sentiment score for the entire dataset if it remained, and thus validated the removal of those observations.

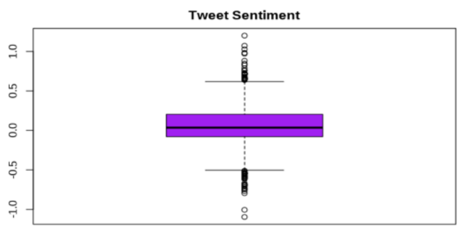

Figure 1: Initial Sentiment Boxplot

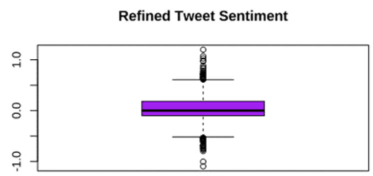

Figure 2: Refined Sentiment Boxplot

Figure1: Initial Sentiment **Figure 2: Refined Sentiment**

Date Analysis
By changing the date variable format, the research team could analyze the data distribution to understand the total number of observations generated daily throughout the collection period. To conduct the distribution analysis, the research team created a histogram to outline the frequency of the variable (Figure 3). The x-axis in the chart represents the calendar day of February 2021. The y-axis contains the number of observations during the collection period.

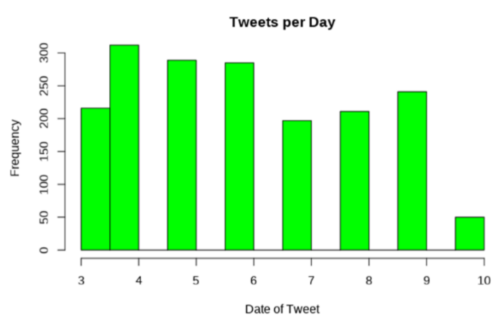

Figure 3: Frequency Histogram

Geographic Sentiment Analysis
On the geographically grouped data, the authors observed a disparity between activity levels on the Twitter platform in the data set. Ideally, the data set would be comprised of a normally distributed frequency between all 50 States. It was observed that the most active State on Twitter represented 255 tweets out of 1791 observations. Inverse to the activity observed in the most active State, there were several States where participation was as low as a single observation. The four most active States during the collection period represents 40.5% of the observations in the dataset (Figure 4). Regardless of the inequity of the participation levels in each State, the conclusion from the research team was that because there were enough States represented in significant enough quantities, the hypothesis could still be tested.

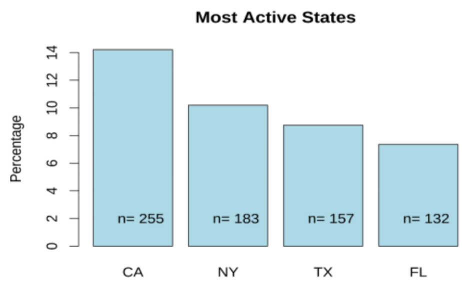

Figure 4: Activity Frequency Chart

We explored the data to understand the sentiment level in each State and concluded that each sentiment mean was likely not representative of the entire State population's *feeling* towards the COVID-19 Vaccine especially because the frequency distribution identified that some sentiment means were derived from tiny amounts of observations. The numeric point chart below (Figure 5) represents the mean sentiment score for each State during the collection period.

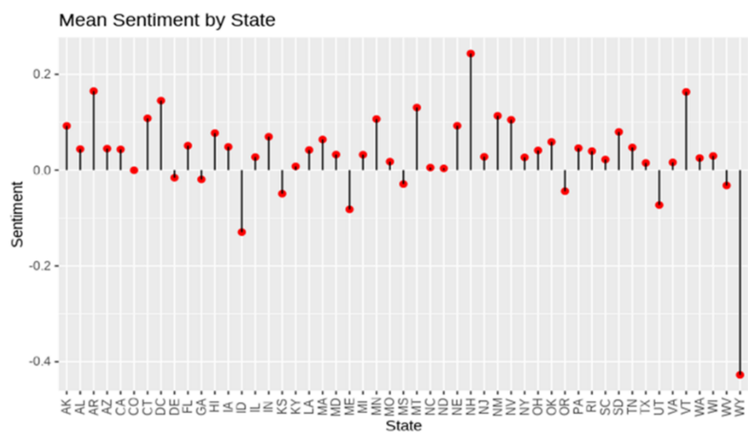

Figure 5: Sentiment Numeric Point Chart

The x-axis in the above chart (figure 5) represents the 50 US States. The y-axis contains the mean sentiment score. It is important to note that the US State of Wyoming on the far right of the chart represents an instance of a mean sentiment score created by only a single observation. However, by looking at the most represented US States of California, Florida, Texas, and New York, the mean sentiment score moves closer to neutral than seen in less represented States in the dataset. Political polarization is generally assumed to occur along geographical parameters within the United States.

Therefore, during the onset of the study, the research team believed that the mean sentiment score would likely reside aligned with the State's political leanings. For that reason, the research team was required to analyze each State's mean sentiment and compare the findings against previous assumptions. The graph below (Figure 6) is a heat map chart illustrating the mean sentiment from all 50 States. Each State's geographic boundaries are filled with a color scale based on the mean sentiment score.

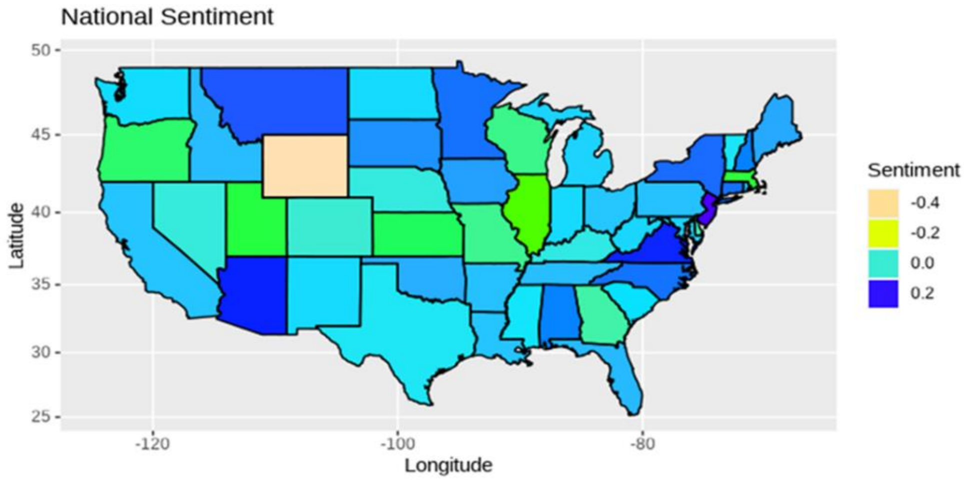

Figure 6: Sentiment Heat Map

Based on the information available, the research team concluded that during the collection period, sentiment means were not directly aligned with the political leanings of the State. The mean sentiment scores varied independently of the political landscape within the United States. Through this chart, Wyoming further solidifies its single observation as an outlier.

Chronological Sentiment Analysis

As the target variable in our test is the sentiment score of each observation, the research needed to understand all factors relevant to the sentiment variable within the dataset. The chart below (Figure 7) provided the first glimpse to the research team of the properties of the sentiment variable. The research team noted that the sentiment variable followed a near-normal distribution with slight negative skewness from 1,791 observations. We concluded that the sentiment variable passed the first assumption for future linear regression testing if warranted because the distribution was nearly normally distributed, with the bulk of the data located close to the mean.

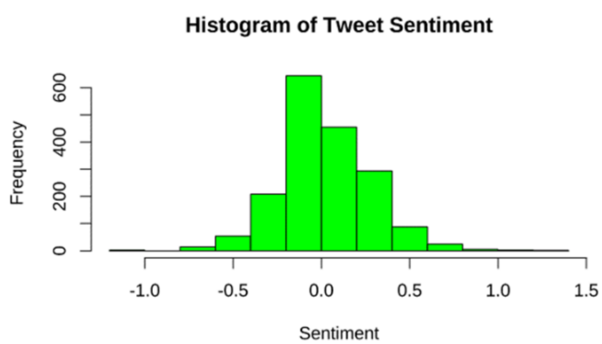

Figure 7: Sentiment Distribution

The following property of the sentiment variable was the movement of the mean throughout the observation period. To accomplish understanding, analysis required grouping the observations by date and extracting the mean for each observation date. The graph below (Figure 8) represents the mean sentiment score from each day of collection. The x-axis outlines the date of observation and the y-axis illustrates the mean sentiment score. From this chart below, the research team identified movement of the mean score throughout the collection period. Additionally, the 8th of February 2021 contained the highest mean sentiment score, where the 10th of February 2021 contained the lowest mean sentiment score from the data set.

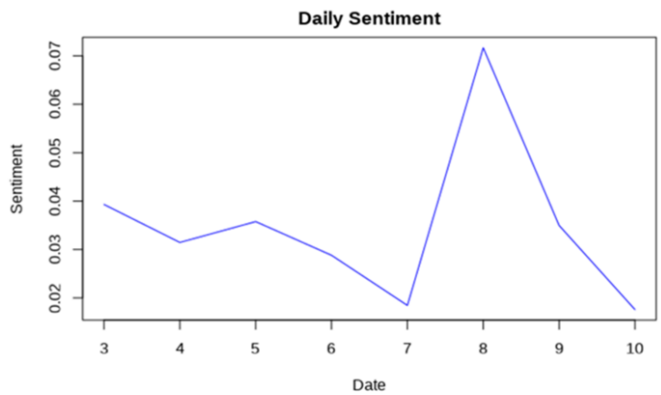

Figure 8: Mean Sentiment Score Line Graph

The tables below (Figure 9) represent the separation of sentiment level per day and the frequency of positive, neutral and negative values appearing across each day of the dataset timeframe. This step provides the separation of values necessary to produce histogram analysis divided into days.

Positive

created_at <dbl>	n <int>
3	107
4	153
5	134
6	135
7	91
8	112
9	112
10	25

Neutral

created_at <dbl>	n <int>
3	33
4	42
5	52
6	36
7	34
8	28
9	45
10	8

Negative

created_at <dbl>	n <int>
3	75
4	117
5	101
6	111
7	71
8	71
9	81
10	17

Figure 9: Sentiment Level Histogram (Positive, Neutral, Negative)

Grouped Sentiment Analysis

During the exploratory analysis phase of the study, the research team sought to identify the distribution of tweets by type. As previously stated, all observations with a sentiment score greater than zero were annotated as positive, observations with sentiment scores less than zero were annotated as negative, and sentiment scores equal to zero were classified as neutral. The bar chart below (Figure 10) displays the frequency of each type of observation throughout the collection period. The x-axis in represents each day during the collection period, and the y-axis provides the total number of observations. Each type of sentiment-grouped observation is represented by a different color outlined in the legend on the right side of the chart.

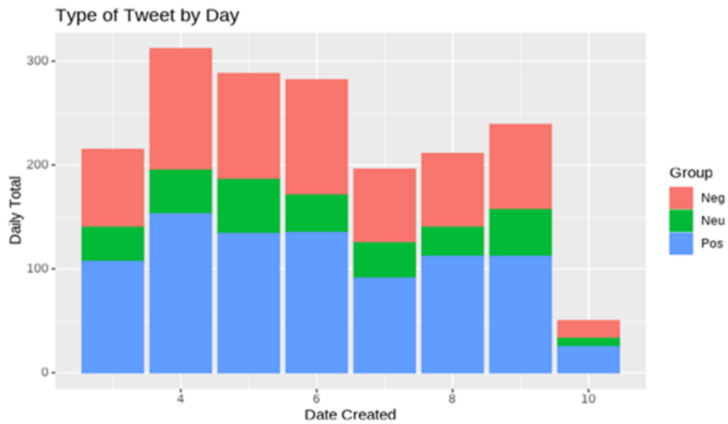

Figure 10: Frequency by Type

Several conclusions can be drawn from the chart above (Figure 10). First, the research team identified that the frequency of the number of observations by date was not normally distributed. As seen through the analysis following the grouping of tweets based on sentiment during the observation period, there are significant differences between the most active days in the data set and the least active days in the data set. The 4th of February 2021 contained the largest number of positive and negative observations during the collection period but contained only a few neutral tweets. The 10th of February 2021 contained the least amount of all three types of sentiment groupings in the data set.

Furthermore, we sought to identify any movement of the percentage of each type of tweet during the observation period. The chart below (Figure 11) illustrates the cumulative relative frequency analysis of each kind of sentiment group throughout the collection period. The x-axis represents the date of observation. The y-axis represents the total percentage of observations. Each sentiment-grouped observation is represented by a different color outlined on the right side of the chart.

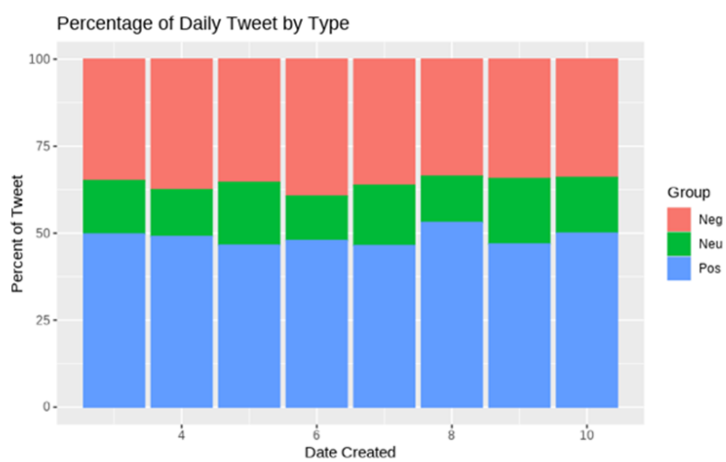

Figure 11: Cumulative Relative Frequency by Type

Based on the cumulative relative frequency of the sentiment types for each date of the observation period, it became clearer that while the mean sentiment score fluctuated throughout the observation period, the type of observation remained relatively consistent.

Issues in Information Systems
Volume 24, Issue 4, pp. 115-129, 2023

Textual Analysis

Textual analysis techniques were employed to investigate variable contributions to sentiment values and frequency of use within the 1,791 observations of Twitter data. Word Cloud analysis was also conducted on each level of sentiment (Figure 12) ranging from below zero (negative sentiment), equal to zero (neutral sentiment), and above zero (positive sentiment). The word clouds below show positive, neutral and negative descriptors displayed from the text variable for the sentiment-grouped observations. While there is a slight variance in each type of word cloud, each one represents a similar pattern found within the comprehensive sentiment analysis.

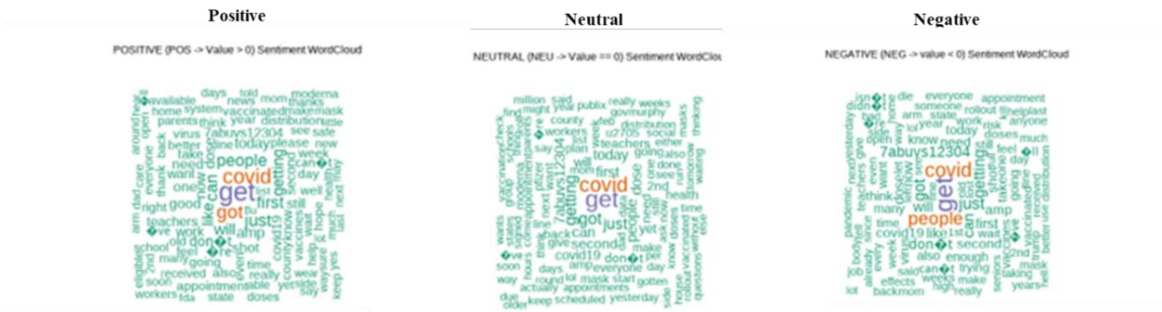

Figure 12: Word Clouds (Positive, Neutral, Negative)

Although there are many common word similarities related to Covid-19, there are some stark differences. Within the Negative Word Cloud on the left, words such as ‘kill’, ‘die’, ‘help’, and ‘pandemic’ tell a story of negative sentiment when communicating on Twitter about the COVID-19 vaccine. As expected, the Neutral sentiment Word Cloud in the center does not reveal any terms which may be considered either negative or positive. Common words of neutral sentiment in the Neutral Word Cloud are: ‘distribution’, ‘waiting’, ‘tomorrow’, ‘appointments’, and ‘scheduled’. For the Positive Word Cloud words such as ‘thanks’, ‘safe’, ‘health’, ‘better’, ‘good’, ‘received’, and ‘want’, relate a message of positivity toward the COVID vaccine. The variety of words utilized within the Twitter data set observations shows a strong mixture of all three sentiment levels.

Variance Testing

After recognizing that the means are numerically different in each of the States, the research team sought to determine if the results of the observations were actual or just due to chance. The test to assess equality of means is a one-way ANOVA test. The data was first subset only to contain the relevant variables. The variables remaining in the dataset are the factors represented by the geographic borders of the State and the sentiment score for each observation associated with the State where the observation occurred. During the ANOVA test, the research team evaluated if the mean sentiment value in California was equal to the mean sentiment value in Delaware, Indiana, Nebraska, and every other State included in the data set. Before running the ANOVA test, the observation containing data from Wyoming was removed as it only represented a single observation and therefore had no variance to be tested.

The purpose of the ANOVA test is to measure variance within the observations. By running a one-way ANOVA test, the research team now accounts for the fluctuation of the sentiment observations in the data set. The null hypothesis for the ANOVA test is that the means are equal in all States, and the alternative hypothesis is that at least one US State has a different mean. During the ANOVA test, the returned p-value was greater than our predetermined confidence level of .05. The p-value for the ANOVA test was .845. Therefore, the research team has concluded there is not sufficient evidence to reject the null hypothesis. Furthermore, because the ANOVA test returned a p-value that was not significant, there is no need for an ad-hoc test to determine which States have significantly different means.

The snapshot pictured below (Figure 13) is the culmination of data synthesis in the form of an analysis variation (ANOVA) test that shows that the P-value is greater than .05, and that there is evidence to accept the Null hypothesis that the state means are equal.

```
            Df Sum Sq Mean Sq F value Pr(>F)
State       49   2.43 0.04965   0.795  0.845
Residuals 1740 108.70 0.06247
```

Figure 13: ANOVA Test Results

Discussion

Data Cleaning and Exploratory Analysis

Data cleaning is necessary for the exploratory phase to ensure downstream conclusions are created without any impacts from data noise which can result from incomplete or faulty information. The first step in the data exploration process was to identify incomplete entries and examine the effects incomplete entries could have on future analysis (Figure 14). During this phase, the research team realized only certain variables needed to be complete to enable proper analysis. The snapshot below (Figure 14) shows a graphical analysis of the completeness of each variable in the dataset across all 31 variables in their original form.

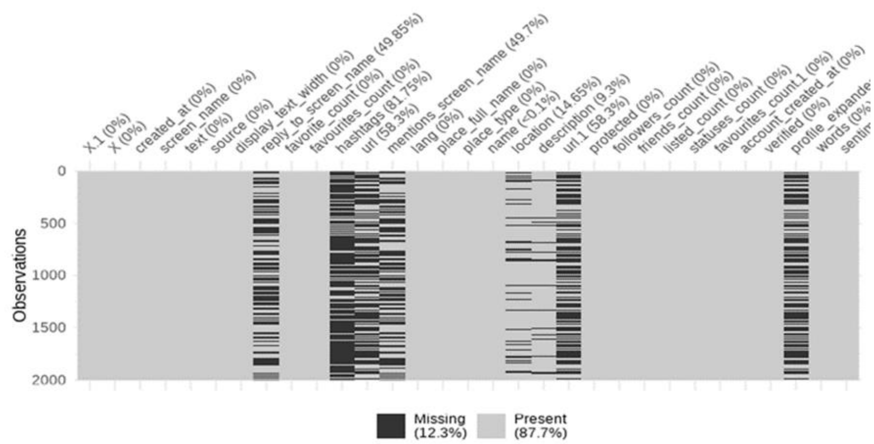

Figure 14: Data Noise

As illustrated in the figure, approximately 12.3% of the data contained blank entries in the observations within the dataset. While blank entries did not immediately concern the research team, further exploratory analysis and manipulation steps were required to create a data set adequate for analysis.

It was important that the research team perform exploratory analysis to understand the type and structure of the dataset. The dataset was comprised of 31 distinct variables in multiple forms such as character, integer, logical, and numeric (Figure 15). Each type of data is represented by a color across all variables class types. The y axis represents each observation within the dataset. The "N/A" or missing dataset values were contained in character class variables, but did not negatively affect the analysis process.

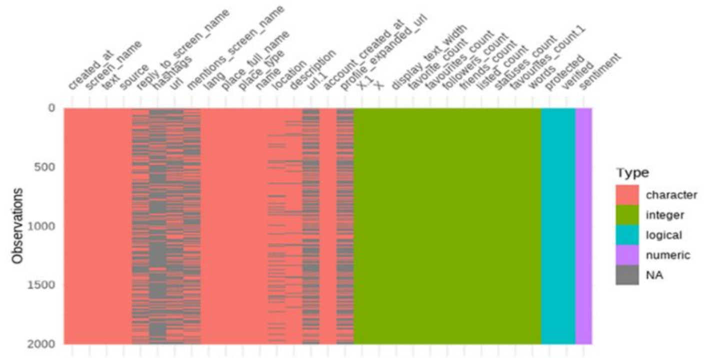

Figure 15: Data Classes

Following structural and completeness exploration, the research team needed to explore the validity of including each variable in its analysis. When initially pulled from Twitter, the dataset contained 31 variables. For the purposes of the study, several variables needed to be excluded from the final dataset. Among the variables removed were the URL, screen name of the author, and observations identified as being created by a bot. Additionally, the research team conducted this exploratory analysis phase to ensure there were no duplicate observations within the dataset. After removing advertisements and observations authored by bots, the original data set of 2,000 observations were trimmed to a total of 1,801 observations. After the cleaning process, the research team moved into the manipulation phase of exploratory analysis.

The dataset contained observations across an eight-day collection period, analysis required manipulating the variables into a form that enables analysis. The year and month values were removed from the variable that annotates the date of creation of the observation. The result was that the new variable annotating the creation date for each observation was now listed as a numeric entry ranging from 3 to 10, representing a calendar date between February 3rd, 2021, and February 10th, 2021.

Next, the research team needed to manipulate the information required for future geographical sentiment analysis. In the original form, the dataset was comprised of several issues that, if ignored, could lead to faulty conclusions. First, the research team identified instances where a name of a US State was located inside the city variable, and the text "USA" was found within the State variable. To rectify the issue, the research team separated the City and State variables to isolate the erroneously collected "USA" entries and created a new data frame that was a subset of the original dataset. The State variable was subsequently reformatted in the new data frame to change entries where two-word US State names lacked separation between words. Once complete, the research team conducted a bind of the isolated entries to the subset of the original data frame. Finally, the research team removed the "City" variable as the future analysis was only geographically focused at the State level.

The result of the process was a consistent variable representing the US State where the observation was created. During this process, the research team also identified the presence of "n/a" values within the dataset in the State variable. These were omitted because they only accounted for seven observations, and exclusion from the data set would not significantly affect the future analysis. After the cleaning phase of the research, the data set was reformatted and trimmed from 1,801 observations to 1,791 observations.

Geographic Sentiment Analysis

The purpose of the study aimed to form a greater understanding of the sentiment toward the COVID-19 vaccine and its distribution throughout the United States. Widely regarded as a divisive topic within

American society, the Twitter data set allows for regional conclusions by sampling the Twitter data. The study conducted its analysis by grouping the Twitter dataset into States, then extrapolating each US State's mean sentiment score. The grouping of observations by State was necessary for testing the hypothesis that the mean sentiment in every State is equal.

Chronological Sentiment Analysis

Nested within the exploratory analysis phase of the study, the research team sought to understand how the data was distributed and understand the movement of sentiment up and down throughout the collection period. The purpose of chronological sentiment analysis was to understand several critical aspects of the dataset. First, the chronological analysis revealed the distribution of tweets arranged by date of creation. The frequency distribution of the number of tweets created by day during the collection period was not normally distributed. Second, the chronological analysis provided insight into further research by observing the movement of the mean sentiment score by date of creation.

Grouping Data by Sentiment

Further refining the dataset into groups of tweets provided the study additional insights into the distribution of data not initially apparent. The study grouped the data based on the sentiment score of each tweet to understand the frequency distribution of each type of tweet. If the sentiment score was greater than 0, it was classified as a positive sentiment tweet denoted by the shorthand Pos. If a tweet sentiment score was less than 0, it was classified as a negative sentiment tweet characterized by the shorthand Neg. When the sentiment score was equal to 0, the tweet was classified as a neutral tweet with the shorthand ‘Neu’.

After the research team understood the frequency of each type of tweet, additional analysis was required to ensure the accuracy of conclusions. Determining the cumulative relative frequency of each type of tweet throughout the collection period provided further insight into the contents of the data set to the research team. The cumulative relative frequency analysis described the percentage that each type of tweet represented in each day’s collection. It was essential for the research team to understand if there were instances where a specific type of tweet significantly increased or decreased throughout the collection period. While the purpose of the study is a hypothesis test of mean sentiment throughout all States, focus was placed throughout the analysis to identify areas that could potentially warrant further exploration.

Textual Analysis

Important to the research team, was achieving and understanding the language used to create a sentiment score. Because a sentiment score is derived from multiple English dictionaries, analysis of the frequency of words used in each type of sentiment-grouped text allowed the research team to peer further in the main ideas, concerns, and sentiment of each type of tweet to identify trends and inform conclusions. While the overall sentiment score informs the opinion of the COVID-19 vaccine, the text analysis illuminated the vernacular most associated within the sentiment groups. Prior to analysis, the research team assumed that the language used in each type of sentiment grouped observation would vary greatly because of the polarizing nature of the topic.

Variance Testing

To determine if the mean sentiment score of all geographic regions was equal, the research team conducted a one-way ANOVA test. The confidence level in the ANOVA test was 95%. The purpose of the ANOVA test is to evaluate and compare the means between multiple samples. During the test, the research team can determine if there is statistical evidence to conclude that the population means are statically different from one another by accounting for the variance within each State's observations. If during the one-way ANOVA test it is found that at least one State's sentiment is significantly different, then the research team will conduct ad-hoc testing to determine which US State's sentiment is different.

Weaknesses and Future Research

During analysis, multiple instances highlighted the need for more depth in the data set. For a more thorough study, the research team concluded that future research should occur with an equal number of observations from each State to understand national sentiment better. Additionally, there is concern among the research team that a shallow data set is not representative of the country as a whole and conclusions only valid for the data set do not adequately capture true National or State sentiment of the COVID-19 vaccine during the collection period. This was not meant to be a generalizable study, but an exploratory study to provide baseline information in a subject area still in its infancy relative to other subject areas in the sentiment analysis discipline.

Although tests conducted in the research failed to prove that the mean sentiment score varied between states, it was apparent that the sentiment values for each observation were not equal. There were instances where the sentiment score was very negative and instances that sentiment scored highly positive. While the purpose of the study was to determine if there was a geographically arranged variability in sentiment, further research should focus on the outlier observations. When observed through averaged sentiment scores, the tone of conversation on Twitter about the COVID-19 Vaccine was largely neutral and sentiment was observed to not be influenced by location, however, it would be valuable to understand if the most negative or most positive observations were geographically arranged.

Conclusion

To conduct this study, the research team gathered information from a raw data set comprised of 2,000 Twitter observations gathered in February of 2021. At the onset of the study, the research team sought to form a greater understanding of sentiment regarding the COVID-19 vaccine within the United States during the observation period. Additionally, the research team tested a hypothesis that the mean sentiment value was equal in all 50 US States. During the study it became apparent to the research team that there is a significant level of variation regarding sentiment in all US States. Analysis revealed that despite a numerical difference in the mean sentiment value, every US State contributed to positive and negative sentiment groups which created large standard deviations of sentiment values when the dataset was grouped by US State. Through analysis, this study failed to find evidence to support the notion that sentiment towards the COVID-19 vaccine of a populace falls within geographic borders. The sample in the study is representative of 1,791 independent emotions and opinions, and regardless of where a tweet was authored, it does not necessarily mean that the tweet will be positive, negative, or neutral towards the COVID-19 vaccine.

References

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R. (2011). Sentiment Analysis of Twitter Data. Department of Computer Science, Columbia University. Retrieved from https://scholar.google.com/scholar_url?url=https://aclanthology.org/W11-0705.pdf&hl=en&sa=X&ei=d36tZOPyCeeK6rQPlLe4qA4&scisig=ABFrs3xgIsHp7r1RrgleUP7Oarxt&oi=scholar

Barney, N. (n.d.). Sentiment Analysis (Opinion Mining). *Tech Target*. Retrieved from https://www.techtarget.com/searchbusinessanalytics/definition/opinion-mining-sentiment-mining

CDC. (2021). COVID Data Tracker Weekly Review. Better, but Not Good Enough. *Centers for Disease Control and Prevention.* Retrieved September 29, 2022 from https://www.cdc.gov/coronavirus /2019-cov/covid-data/covidview/past-reports/02122021.html

CDC. (n.d.). CDC Museum COVID-19 Timeline. *Centers for Disease Control and Prevention*. Retrieved September 29, 2022 from https://www.cdc.gov/museum/timeline/covid19.html#:~:text=December%2012%2C%202019,respond%20well %20to%20standard%20treatments

Feuerriegel, S. & Proellochs, N. (2021). Sentiment Analysis Vignette. *R-Project*. Retrieved from https://cran.r-project.org/web/packages/SentimentAnalysis/vignettes/SentimentAnalysis.html #built-in-dictionaries

Gupta, S. (2018). Sentiment Analysis: Concept, Analysis and Applications. *Towards Data Science*. Retrieved from https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17

Hill, C. (n.d.). Sentiment Analysis (Lexicons). *RPubs by RStudio*. Retrieved from https://rpubs.com/chelseyhill/676279

Pang, B. & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval*: Vol. 2: No. 1–2, pp 1-135. http://dx.doi.org/10.1561/1500000011

Proellochs, N. & Feuerriegel S. (2021) *Package 'Sentiment Analysis'*. Retrieved from https://cran.r-project.org/web/packages/SentimentAnalysis/SentimentAnalysis.pdf

Rahman, M. M., Ali, G. M. N., Li, X. J., Samuel, J., Paul, K. C., Chong, P. H., & Yakubov, M. (2021). Socioeconomic factors analysis for COVID-19 US reopening sentiment with Twitter and census data. Heliyon (ScienceDirect by Elsevier), e06200

Samuel, J., Rahman, M., Ali, Nawaz G. G. Md., Samuel, Y., Pelaez, A., Chong, P. H. J. and Yakubov, M (2020) "Feeling Positive About Reopening? New Normal Scenarios From COVID-19 US Reopen Sentiment Analytics," in IEEE Access, vol. 8, pp. 142173-142190, 2020, doi: 10.1109/ACCESS.2020.3013933. https://ieeexplore.ieee.org/document/9154672

Samuel, J. (2017). Information token driven machine learning for electronic markets: Performance effects in behavioral financial big data analytics. *JISTEM-Journal of Information Systems and Technology Management*, 14(3), 371-383

Silge, J. & Robinson, D. (n.d.). Welcome to Text Mining with R. *Tidy Text Mining*. Retrieved from https://www.tidytextmining.com/sentiment.html

The Washington Post. (2022). Coronavirus in the U.S.: Latest Map and Case Count. *The Washington Post*. Retrieved October 5, 2022 from https://www.nytimes.com/interactive/2021/us/covid-cases.html