Oviya Adhan
Clarissa Solis
Robin Zhao

Gauging New York Mental Health Through COVID-19 Tweets

1. ABSTRACT

COVID-19 has been, and still is, making its way around the world plaguing millions of people. When COVID-19 made its way to the United States, the first state to see catastrophic infection rates and deaths was New York. As a result, many have been sharing their thoughts, grievances, and worries surrounding the pandemic on social media, and in this case, through Twitter. With this, our group wanted to see how these depressive and anxiety ridden thoughts were related to COVID-19 infection numbers in New York. Throughout our study, we utilized several different methods in order to gauge how the mental health of New Yorkers was affected from March 2020 to April 2021. In brief, first we wanted to demonstrate the validity of Zipf's law within our corpus, which through the analysis, we found that the law held for our corpus. Next, we wanted to compare the variation between sentiment of tweets and the occurrences of depression-related keywords to the variation in new COVID-19 cases in New York State. With our analysis, we found that keyword occurrences and sentiments move together, however, they do not appear to be positively correlated with the increase of COVID-19 case numbers, but are surprisingly negatively correlated. As a result, our hypothesis was not supported and does not show evidence that depression and anxiety related tweets and negative sentiment are positively correlated with the increase of COVID-19 case numbers.

2. INTRODUCTION

2.1. Scientific Question and Motivation

In regards to our rationale behind our analysis, we chose to study this because it is well-known that people's mental health quality declined during the COVID-19 pandemic and wanted to see how it is reflected in tweets, especially because we have experienced it in real time.

2.2. Literature Review

Ever since the immediate isolation that came along with the start of the pandemic, mental health has been closely linked with discussions surrounding the topic of COVID-19. Even the Center of Disease Control (CDC) has created a separate webpage to address mental health risks as a result of the pandemic. This has resulted in a wide range of studies that have captured the relationship between mental health and COVID-19. Ones that capture the dicussion's presence on social media are fewer, but the topic in relation to social media is still commonly researched. One of which we found that is similar to the purpose of this study is an article published in the Health Data Science journal in February 2022 by Zhang et al. from the University of Rochester. The article is titled "The COVID-19 Pandemic and Mental Health Concerns on Twitter in the United States".

This study used a Twitter streaming API to collect COVID-19 related tweets from 03/05/2020 - 01/31/2021. From that they kept tweets that contained the keywords 'depress', 'failure', and 'hopeless'. The research team conducted topic modeling using the Latent Dirichlet Allocation (LDA) model to track Twitter users' discussion surrounding mental health in relation to COVID-19. This was followed by deducing the demographic composition of these users using deep learning algorithms, specifically the algorithm used a facial detection API called Face++ to detect age and gender, and a race/ethnicity predictor package called Ethnicolr to detect race. Through the topic modeling, they found a mild positive correlation of 0.1196 between the number of tweets related to mental health and COVID-19 cases in the US, and they found 'stay-at-home', 'death poll', and 'politics and policy' were the most commonly mentioned topics. Through deep learning, they found that White Male users between the ages of 30-49 were most likely to express mental health concerns through tweets.

After reviewing their research, we found certain important gaps in the research. While the topic modeling techniques appear sound, the deep learning algorithms used appear to introduce bias into the study. Both the Face++ API and the Ethniclr package use facial recognition

technology, which at this stage in its technology, is not the most accurate way to collect data. Face++ is meant to detect age and gender. However, there is no proper way to detect these since age is presented differently among different individuals and gender is not a binary variable and should not be treated as such especially in a study that was conducted in 2021. Ethniclr is meant to detect race, but implying that there are generalizable physical features of different races is harmful and often times inaccurate due to the diversity of appearances within racial groups as well as people who are of mixed race. Additionally, the research team used the sample demographics without comparing them to the Twitter user base demographics. Therefore, these percentage breakdowns may simply reflect Twitter's overall user base rather than for this sample in particular.

For the issues mentioned above, the second half of the study concerning the demographics cannot be seen as viable information due to the bias introduced in several areas of that half. However, the publication's work with topic modeling gives us more insight into the particular topics that were discussed in relation in COVID-19 and mental health. The correlation found at 0.1196 shows us a positive correlation between occurrences of mental-health related tweets and COVID-19 cases in the US, but also does not seem to be high enough to provide strong evidence for their hypothesis.

2.3. Hypothesis

With this corpus analysis, we hypothesize that the number of depression and anxiety related tweets and negative sentiment are positively correlated with increasing COVID-19 cases.

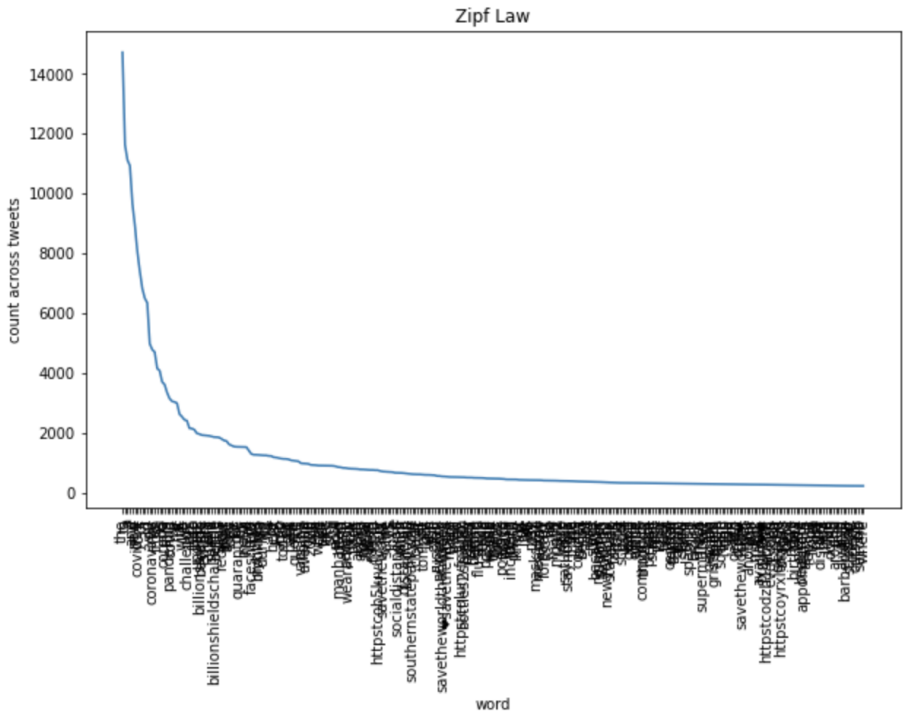
3. METHODS AND RESULTS

The original dataset is the COVID-19 Tweets Dataset cited from IEEE DataPort (Lamsal, 2022). This dataset includes CSV files that contain IDs and sentiment scores of the tweets related to the COVID-19 pandemic. The real-time Twitter feed is monitored for coronavirus-related

tweets using 90+ different keywords and hashtags that are commonly used while referencing the pandemic. The tweets collected range from 2020/3-2021/7.

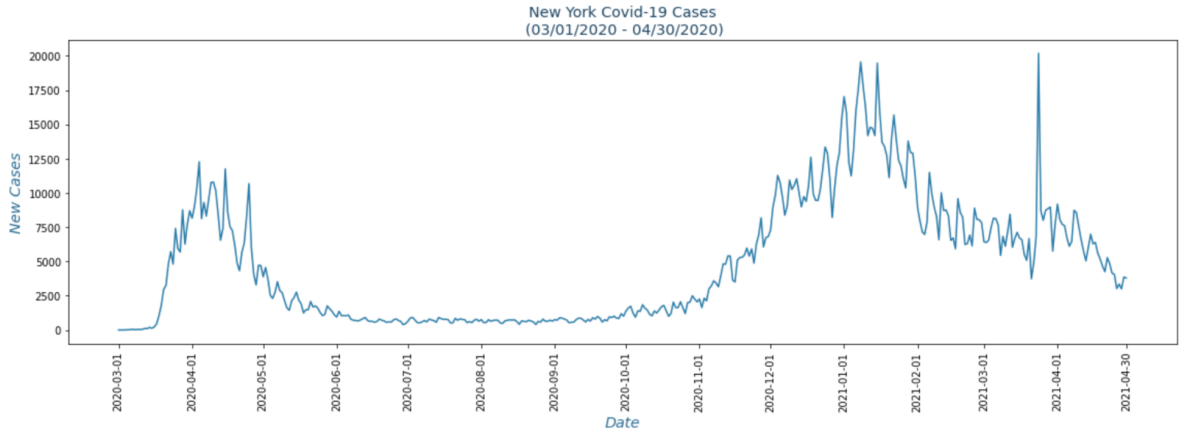
To obtain the actual text of the tweets, we first use Twitter API to request tweet information through its ID. This process is known as hydration. The result of the tweet hydration is a pandas dataframe with columns date (date of the tweet), text, sentiment, retweet count, country, and city (tweet's location). Since we decided to focus on tweets from New York State for our analysis, we filtered the data frame to only contain tweets in which their city contains "NY". The initial data frame contains 24401 rows, which corresponds to 24401 tweets.

The first part of our analysis is to demonstrate the validity of Zipf's Law in our corpus. Zipf's law is a statistical distribution in certain data sets, such as words in a linguistic corpus, and it states that the frequencies of certain words are inversely proportional to their ranks. Each individual tweet first goes through the preprocessing steps including tokenization, conversion to lowercase, removing punctuations, and lemmatization. The list of punctuations is based on punctuation from the string class, and lemmatization is performed using "wordnet" from the NLTK library. The reason why we choose to use lemmatization over stemming is because the corpus is not very big, and the results of lemmatization are more easy to interpret. Next, all preprocessed tweets are concatenated into a string, and Counter class is used to identify the unique words with the number of their occurrences. The result is then sorted by the number of occurrences descending, and a line plot is generated with x being each unique word and y being its occurrences.



From the graph, we can clearly see that the relationship between unique words and occurrences follow a Pareto distribution. This demonstrates that Zipf's Law holds for our corpus.

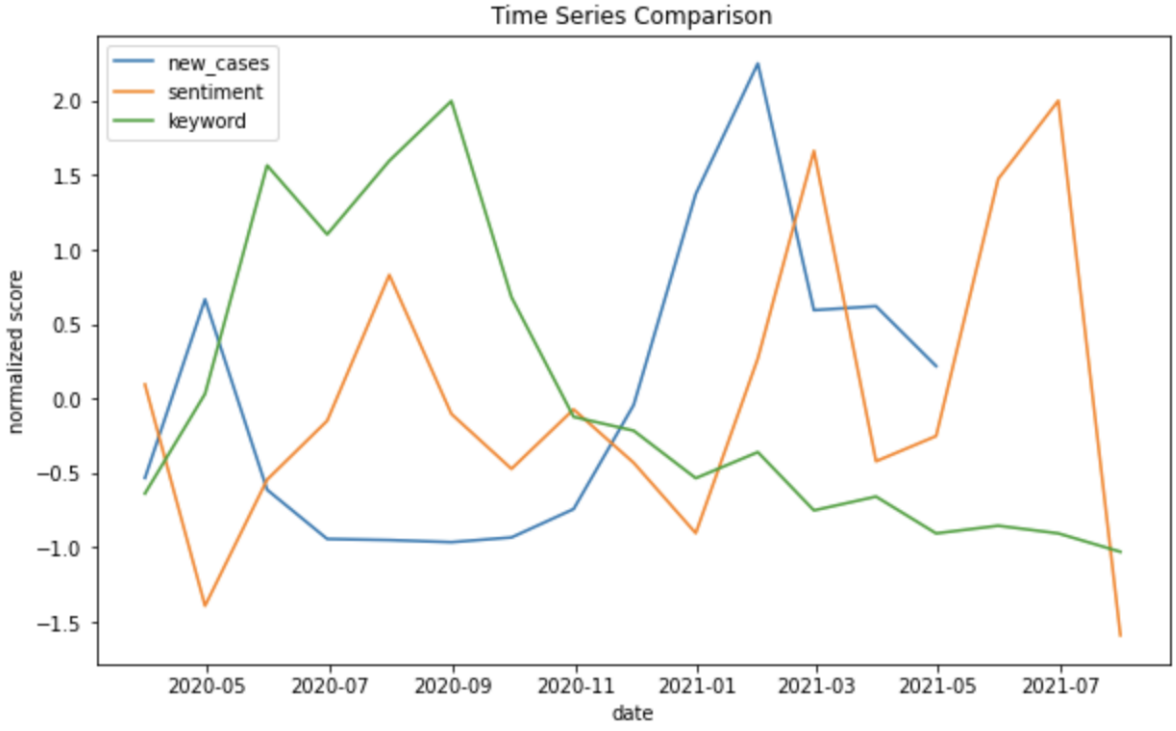
The next part of our analysis aims to compare the variation in sentiment of tweets and the occurrences of depression-related keywords to the variation in new COVID-19 cases in New York State. Still using the tweet dataframe we preprocessed earlier, we further removed stop words using the stopword list from the NLTK library. Besides the tweet dataframe, we also used a COVID dataset (NYC Open Data, 2022) to obtain the number of new COVID-19 cases in New York state per day. This COVID dataset is preprocessed by casting columns to appropriate types and narrowing down the dataset to only NY states and data ranging from 2/28/2020 to 5/01/2021. An overview graph that shows the time series of new cases by date is the following.



To identify depression related sentiment in the tweet dataset, we curated 170 depression related keywords, based on the information from prestigious psychology institute websites like the Berkeley wellbeing institute and Healthline. Some representative keywords include: 'mental', 'anxiety', 'depression', 'worried', 'anxious', 'scared', 'depressed', 'lonely'.

In order to see how much the corpus contains the curated keywords, a lexical dispersion plot is generated with the first ten keywords, which measures how frequently a word appears across parts of a corpus. We can see that the keyword occurrences are not too sparse given the large amount of tweets in the corpus. This can provide some credibility to the final result of our study.

From the previous preprocessed tweet dataset, tweets that contain depression related keywords are extracted, and these tweets are aggregated to the monthly level. Monthly sentiment score is calculated by taking the mean of all extracted tweets sentiment scores in a month, and monthly keyword occurrence count is calculated by the number of depression-related tweets in a month. To compare the monthly sentiment scores and monthly keyword occurrences to the monthly new COVID-19 cases, these three variables are normalized using z score, which is measured in terms of standard deviations from the mean. The final time series graph that compares sentiment scores, depression keyword occurrences to the COVID-19 new cases numbers is shown below.



From the graph, we can see that the time series of sentiment and keyword follow a similar pattern: an increase in keyword occurrences is often followed by an increase in sentiment score. This implies that our depression keyword analysis has some degree of accuracy. Furthermore, the keyword occurrences are surprisingly negatively correlated with the case number. Lag analysis shows that the correlation between case number and keyword occurrences lagged by one week is -0.77, which is a moderately strong negative correlation. This contradicts

with our initial hypothesis, and does not provide evidence that Instances of depression and anxiety related tweets and negative sentiment are positively correlated with increasing COVID-19 cases.

4. DISCUSSION

4.1. Reiterate Conclusions

Through conducting this study, we found that from March 2020 to April 2021, keyword occurrences and sentiments in mental-health related tweets follow a similar pattern. Both keyword occurrences and sentiment did not, however, appear to be positively correlated with the increase in case numbers. For that reason, we reject our hypothesis and we conclude that there is no indicated relationship between the occurrence of mental-health related tweets and COVID-19 case numbers.

4.2. Connect to previous literature

Contrary to Zhang et al., we found a negative correlation between tweets mentioning the keywords and COVID-19 cases. It is important to point out the differences between our studies. Our study covers only New York state over the course of a year and a month from March 2020 to April 2021, while Zhang et al. covered tweets in the US from March 2020 to January 2021. The geographical differences in our populations may account for our varying conclusions. Our correlation -0.77 and their correlation of 0.119 are far off from one another and together suggest contradictory results. While we rejected our hypothesis, Zhang et al. accepted their hypothesis.

4.3. Limitations and future directions

Our study has two limitations. The primary limitation of this study is that we are limited to tweets in English due to the language limitations of NLP packages especially with stopwords, tokenizers, and lemmitizers. With the pandemic, it was clear that underrepresented communities, often BIPOC and immigrant communities who are more likely to speak different languages, were

disproportionately affected by the pandemic. Our study cannot reflect this reality due to being limited to English tweets. Another limitation is that our study is limited to Twitter, so our conclusions can not be generalized to all of social media, which would be more useful.

In the future, this study can be expanded to include a more in depth analysis of these tweets. Similar to Zhang et al., we can look into demographics, but we would use a less biased way of gathering this data. Additionally this could be expanded to gather other tweets from these users to have a more thorough understanding of the general topics that these users tweet about and see how much weight the COVID-19 mental health related tweets have compared to the rest of their tweets. If we had more time for a more thorough project, we would have ideally used Reinforcement Learning to create a model to see if we can predict events, rather than just analyze the effects of an event, based on tweets.

References

Coping with Stress. (n.d.). Stress. https://www.cdc.gov/mentalhealth/stress-coping/cope-with-stress/index.html

COVID-19 Daily Counts of Cases, Hospitalizations, and Deaths | NYC Open Data. (2022, May 13). New York State. https://data.cityofnewyork.us/Health/COVID-19-Daily-Counts-of-Cases-Hospitalizations-an/rc75-m7u3

Lamsal, Rabindra. "Coronavirus (COVID-19) Tweets Dataset." *IEEE DataPort*, IEEE, 12 May 2022, https://ieee-dataport.org/open-access/coronavirus-covid-19-tweets-dataset.

Negative Emotions: List & 158 Examples (+ PDF). (n.d.). The Berkeley Well-Being Institute. https://www.berkeleywellbeing.com/negative-emotions.html

Raypole, C. (2019, September 10). *Big Feels and How to Talk About Them*. Healthline. https://www.healthline.com/health/list-of-emotions#fear

Zhang, S. (2022, February 17). *The COVID-19 Pandemic and Mental Health Concerns on Twitter in the United States*. SPJ. https://spj.sciencemag.org/journals/hds/2022/9758408/