Analyzing Twitter, Reddit and 4chan Trends on Covid-19

Umesh Kumar Gandla
ugandla1@binghamton.edu
Binghamton University
Binghamton, NY, USA

Goutham Metta
gmetta1@binghamton.edu
Binghamton University
Binghamton, NY, USA

Manikya Pandey
mpandey2@binghamton.edu
Binghamton University
Binghamton, NY, USA

Harsha Vardhan Goud Maragoni
hmarago1@binghamton.edu
Binghamton University
Binghamton, NY, USA

Ashika Yellamelli
ayellam1@binghamton.edu
Binghamton University
Binghamton, NY, USA

ABSTRACT

The key to controlling the COVID-19 pandemic is having an understanding of how people respond to COVID-19 vaccines. Examining public opinion is necessary in order to comprehend the reaction. In this project, we are collecting data from social networking sites like Twitter, Reddit, and 4chan. These sites provide discussion boards for exchanging information on a range of global problems, like COVID-19. The purpose of this study was to examine public perceptions and responses to the COVID-19 pandemic using data from social media sites like Twitter and Reddit, as well as to investigate politics using the Reddit data set. Traditional surveys only cover a tiny subset of health-related subjects, collect data on a small scale, and are costly and time-consuming. In order to gauge public opinion on the Covid-19 vaccination and collect textual data from social networking sites like Twitter, Reddit, and 4chan, we are conducting a textual information collection.

In this project, we want to analyze the Covid - 19 trends on Reddit, Twitter, and 4chan. we are using Twitter sample stream API, Reddit API, and 4chan API for the data collection process which provides access to daily posts on these social media sites.

1 INTRODUCTION

A particular strain of flu first appeared in Wuhan, China, in January 2020, and it quickly spread throughout the world via tourists. In March 2020, the World Health Organization classified it as a pandemic. Up until mid-May 2021, there have been more than 158 million confirmed instances of Covid-19 worldwide, and the number of fatalities has surpassed 3.2 million. More than 32 million cases and 580,000 fatalities were recorded in the United States. Because of the lockdowns brought on by this pandemic, more people are now using social media sites like Twitter and Reddit to share news and express their thoughts. These forums have been used by both individuals and government representatives to frequently discuss COVID-19-related policies and news.

Social media, according to research trends, will be crucial for promoting public health. Social media can assist in public health surveillance by facilitating real-time communication at a low cost and monitoring public reaction to health crises. In this case, Twitter and Reddit emerge as crucial information sources that mostly contain personal opinions and responses to those viewpoints rather than posts from carefully constructed content developers.

As more people are getting vaccinated, the COVID-19 dataset can be used to assess how the general population feels about vaccination, how that feeling has changed over time, and what effects vaccination has had. Herd immunity is negatively impacted by low vaccination adoption. Therefore, a significant success element for creating non-pharmaceutical therapies and managing COVID-19 is the public's response to COVID-19 vaccinations. In order to study the trend in user sentiments linked to the Covid-19 Vaccine, this project will extract, examine, and explore postings from Twitter and Reddit that are related to the topic.

2 BACKGROUND

When the World Health Organization declared COVID-19 a pandemic on March 11, 2020, the illness had an unprecedented worldwide impact. During a crisis, social media sites like Twitter, 4chan, Reddit, and others may be useful tools for situational awareness by keeping an eye on public attitudes and behavior. A better understanding of the public's attitudes and actions during the COVID-19 problem may be achieved by using the information acquired to support communication and health promotion activities. This study sought to compare popular perceptions of the epidemic.

3 DATA COLLECTION

3.1 Twitter's Sample Stream API

Twitter provides API endpoints to extract and monitor the constantly evolving trends across a wide range of topics discussed on the platform. We used the Twitter Sample Stream API to extract the tweets from the platform. The Twitter API allows you to stream public Tweets from the platform in real time so that you can display basic metrics about them on some data points like Tweets, Users, and IDs.

we collected data from 1 million tweet posts from Nov 1, 2022, to Nov 14, 2022, i.e 14 days of the data extracted for the Twitter data set. The extracted data is stored in MongoDB in the form of a CSV file.

3.2 Reddit API

We use Reddit API which allows you to extract data, or post to Reddit using a web application or your preferred programming language. We used it to extract the posts, and comments from specific subreddits. Reddit has vast data to analyze ongoing trends on covid for this project.

we had extracted approximately 3 million posts from Nov 1, 2022, to Nov 14, 2022, i.e 14 days of data were extracted from Reddit. The extracted data is stored in MongoDB in the form of a CSV file.

3.3 4Chan API

We used the 4Chan catalog API to extract data from 4chan-specific boards and filtered data related to covid-19. we collected data points like id, post, and date.

We extracted 225K posts from Nov 1, 2022, to Nov 14, 2022, i.e 14 days of data was extracted from 4chan. The extracted data is stored in MongoDB in the form of a CSV file.

4 PROJECT FLOW

We will retrieve real-world data from Twitter Sample stream API, Reddit API, and 4chan API using Python modules. Then Store data in MongoDB and clean the data using python techniques. We use a Machine learning algorithm to perform measurements and analysis of the data. We also did sentiment and Subjective analysis on the posts extracted from all three data sources. Using Python modules, we categorize the data and map the job trends by geography and industry.

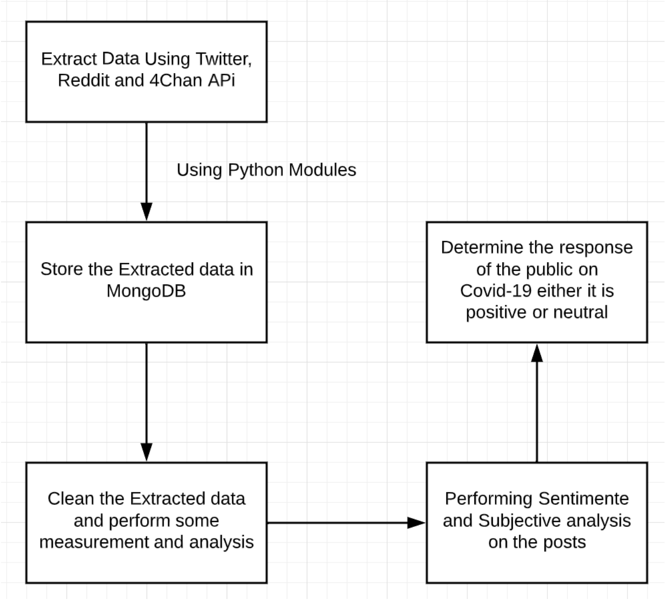

Extract Data Using Twitter, Reddit and 4Chan APi
Using Python Modules
Store the Extracted data in MongoDB
Clean the Extracted data and perform some measurement and analysis
Performing Sentimente and Subjective analysis on the posts
Determine the response of the public on Covid-19 either it is positive or neutral

Figure 1: DATA FLOW DIAGRAM.

5 DATA VISUALIZATION

5.1 Sentimental Analysis on Twitter, Reddit, and 4Chan

The below linear graph shows the sentimental analysis of posts or tweets from Twitter, Reddit, and 4chan. In the below graph 0 means, a post is neutral, 1 means the post is positive and -1 means a post is negative. The graph plots all posts collected as positive,neutral,and negative.

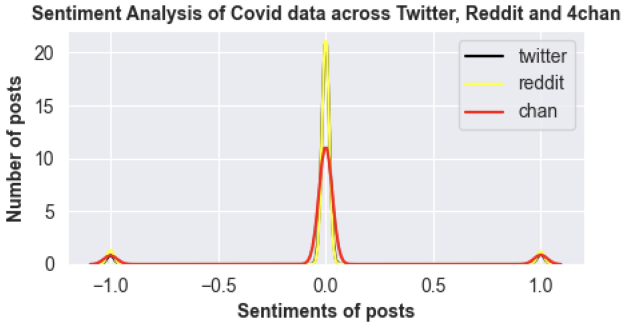


Figure 2: Sentiment Analysis on all three data sources.

5.2 Sentimental Analysis on Reddit dataset

The below graph shows the sentiment scores of posts from Reddit and gives scores as positive, negative, and neutral scores. The graph plots positive, neutral, and negative scores of all posts in the Reddit dataset as subplots.

Sentiment Analysis is a use case of Natural Languagae Processing and comes under the category of text classification. To put it simply,sentiment analysis involves classifying a text into various sentiments such as positive or negative, happy, Sad or Neutral. Thus, the ultimate goal of sentiment analysis is to decipher the underlying mood, emotion or sentiment of a text.

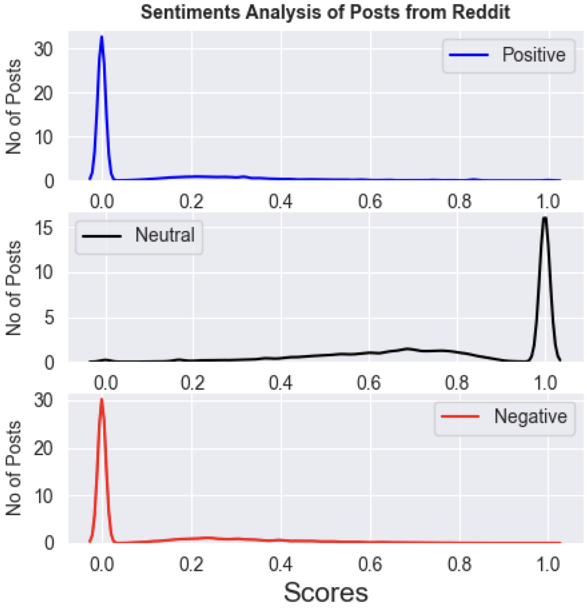


Figure 3: Sentiment Analysis on all Reddit posts.

5.3 Sentimental Analysis on Twitter dataset

The graph below displays the positive, negative, and neutral sentiment scores for tweets extracted from twitter. Positive, neutral, and negative scores for each tweet in the Twitter dataset are plotted separately on the graph.

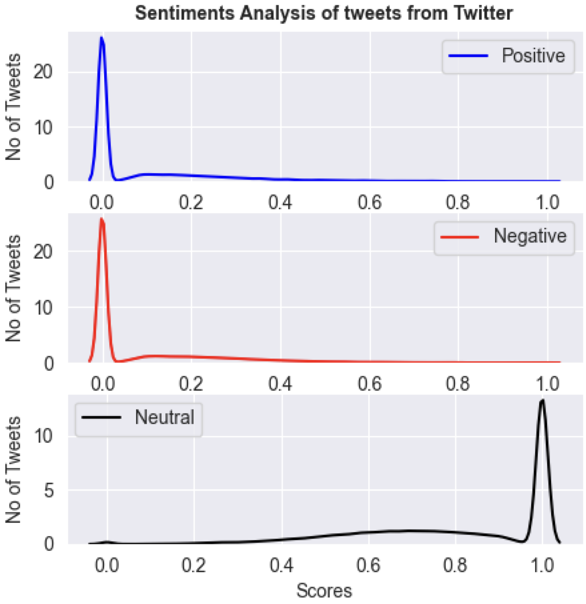


Figure 4: Sentiment Analysis on all Twitter tweets.

5.4 Sentimental Analysis on 4Chan dataset

The graph below displays the sentiment scores for posts taken from 4chan and assigns each post a positive, negative, or neutral score. Positive, neutral, and negative scores for each post in the 4chan dataset are subplots on the graph.

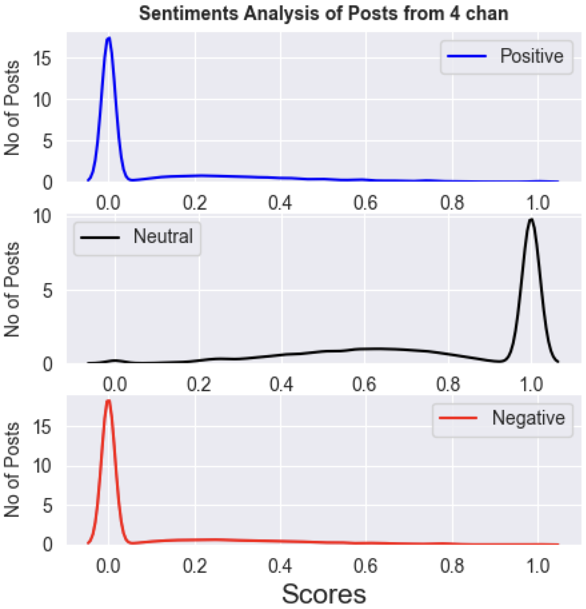


Figure 5: Sentiment Analysis on all 4Chan posts.

5.5 Analysis on most repeated words in Twitter and Reddit datasets

The graph displays a bar plot of the top 10 most often used words from postings in Twitter and Reddit databases.

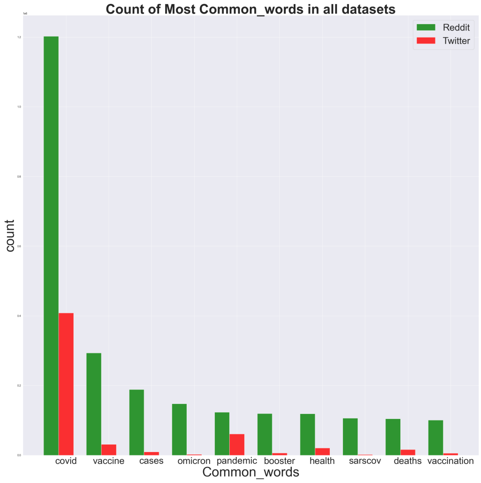


Figure 6: Analysis of most repeated words.

5.6 Illustration of Number of tweets extracted from Twitter

The line graph below shows how many tweets were taken from the Twitter sample stream throughout the course of time. The number of tweets received from November 1 through November 14 is represented on the y-axis, and the x-axis is daily binned data.

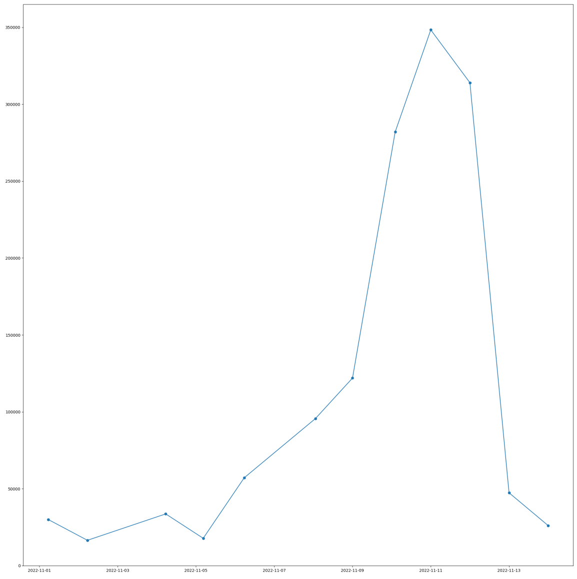


Figure 7: Tweets Extracted Daily.

5.7 Illustration of Number of posts extracted from Subreddit Politics

The line graph below displays the number of submissions every day on the r/politics subreddit from November 4 to November 14, 2022. The data is displayed on the x-axis, counted on the y-axis, and binned hourly on the x-axis.

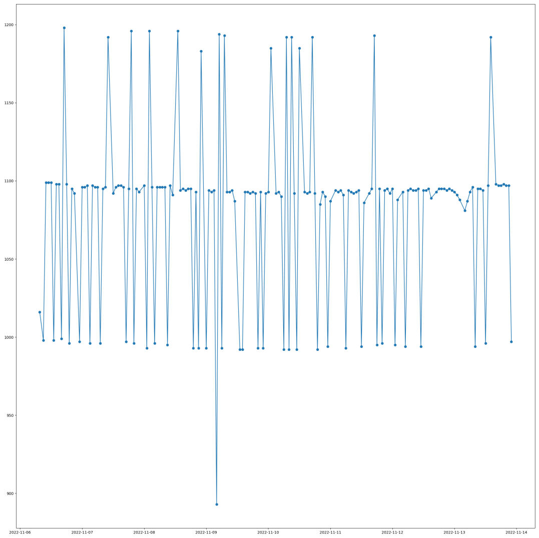
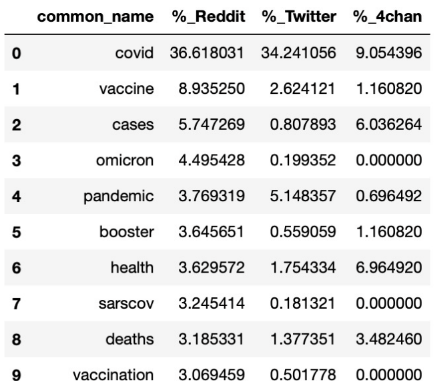


Figure 8: Posts Extracted Daily from r/politics.

5.8 Data comparison utilizing popular search terms from Twitter, Reddit, and 4chan

The percentage of popular words posted on Twitter, Reddit, and 4chan is displayed in the table below. It demonstrates the language usage patterns for the COVID-19 pandemic on social media sites.



	common_name	%_Reddit	%_Twitter	%_4chan
0	covid	36.618031	34.241056	9.054396
1	vaccine	8.935250	2.624121	1.160820
2	cases	5.747269	0.807893	6.036264
3	omicron	4.495428	0.199352	0.000000
4	pandemic	3.769319	5.148357	0.696492
5	booster	3.645651	0.559059	1.160820
6	health	3.629572	1.754334	6.964920
7	sarscov	3.245414	0.181321	0.000000
8	deaths	3.185331	1.377351	3.482460
9	vaccination	3.069459	0.501778	0.000000

Figure 9: Comparison of words.

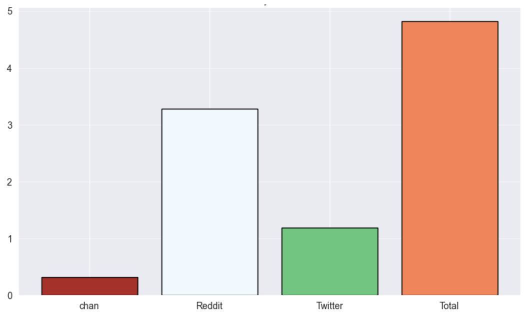
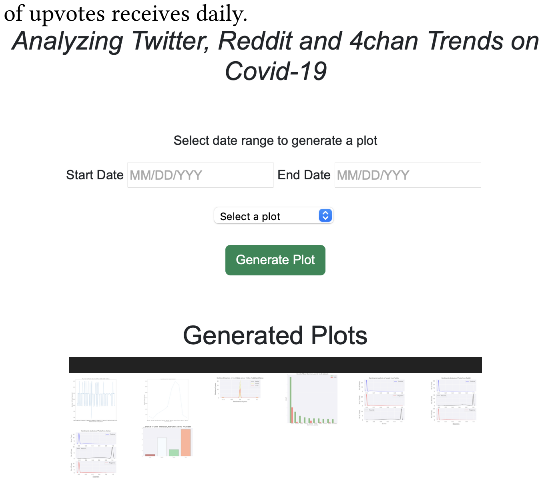


Figure 10: Total data from Twitter, Reddit, and 4Chan.

6 DASHBOARD IMPLEMENTATION

We ultimately created a web application for the dashboard using Python flask and HTML that produces the requested plot between two dates that are entered in MM-DD-YYYY format.

In this, a project we developed a web-based application that takes the start date and end dates as input and generates the plots from the MongoDB. In this application, we generated two plots. One displays the average subjective analysis of the posts extracted from Twitter, Reddit, and 4chan. In the second plot, we generated the number of upvotes received on Reddit on a daily basis and the no of upvotes receives daily.



Analyzing Twitter, Reddit and 4chan Trends on Covid-19

Select date range to generate a plot

Start Date MM/DD/YYY End Date MM/DD/YYY

Generated Plots

Figure 11: User Interactive UI .

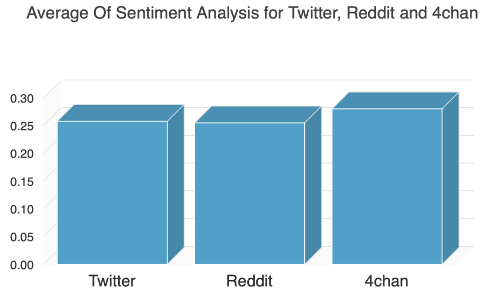


Figure 12: Average Subjective Analysis on Twitter, Reddit, and 4Chan.

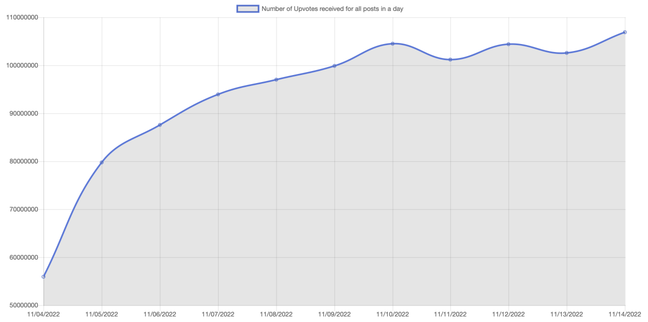


Figure 13:Number of Upvotes received.

Figure 11 shows the home page of our flask web application, where we may enter dates in the format of MM-DD-YYY to produce the above two plots. The dates run from 11-1 to 11-14. We used vscode to run our web application on the virtual machine and it runs on the local host.

6.1 Research Question

Determine whether the public responses to Covid-19 are positive or negative and whether is it similar or not on all 3 social media sites. By doing a subjective Analysis of three data sources and generating a bar plot like figure 11 we can tell that most of the posts or tweets are positive and public response is good towards the covid-19. In the second plot, we generated the number of upvotes received on Reddit daily, and the line graph increases daily from that we can confirm that public response is increasing and it's more positive than negative.

6.2 BUG-Report

During the creation of Subjective analysis and Upvotes of Reddit plots, as well as during the implementation of the dashboard, we had not found any issues. Reddit's dashboard takes some time to plot because the dataset is so large.

7 CONCLUSION

Our study's findings concur with those found in the literature about the epidemic's negative impact on people's emotions and psychological states. The research's data sets further the amount of empirical evidence demonstrating the pandemic's multiple negative consequences on society. In time the opinion of the pandemic or response of the public towards covid has gradually changed and the public is responding more positively.

Additionally, it demonstrates how to apply ML methods in conjunction with statistical modeling and inference to make better use of real-world data. Researchers will be able to conduct a variety of studies using this open dataset, including ones on the psychological and emotional responses to social distance measures, the identification of false information sources, and the stratified measurement of attitude toward the epidemic in almost real-time.

8 REFERENCES

1. https://developer.twitter.com/en/docs/twitter-api/tweets/volume-streams/api-reference/get-tweets-sample-stream

2. https://www.reddit.com/dev/api/

3. Beraldo, D. 2017. Contentious Branding: Reassembling Social Movements Through Digital Mediators. PhD Dissertation. Faculty of Social and Behavioural Sciences, Amsterdam Institute for Social Science Research, Amsterdam.

4.De Zeeuw, D.; and Tuters, M. 2020. Teh Internet Is Serious Business: On the Deep Vernacular Web Imaginary. Cultural Politics 16(2).

5.https://www.openicpsr.org/openicpsr/project/120321/ version/V12/view;jsessionid=30EC3A1F5CFF2E59B3A1E38BA8B94044

6.https://ieee-dataport.org/open-access/ coronavirus-covid-19-tweets-dataset

7.https://www.researchgate.net/publication/360641108 $_{S}entiment_{A}nalysis_{o}f_{C}ovid$- $related_{R}eddits$

8.http://cs229.stanford.edu/proj2016/report/AcostaIlcusWegrzynski-Predicting

9. Hutto, C.J. Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014

10. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7505605/

11.https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0249037