Exploring the Sentiment of COVID-19 Related Tweets: A Study of Twitter Discourse During the Pandemic. (1468 words)

Divya Dhaipullay, MS in Data Science, 2nd Semester, Indiana University

1. Introduction

The COVID-19 pandemic has caused unprecedented challenges worldwide, affecting various aspects of daily life. Social media platforms, particularly Twitter, have played a significant role in disseminating information and shaping public opinion [1]. Several studies have investigated the use of Twitter data to track and predict the spread of COVID-19. For instance, Li et al. [2] analyzed Twitter data to predict COVID-19 cases at the county level in the United States. Similarly, Singh et al. [3] explored the use of Twitter data to monitor and understand the public's perception and sentiment towards the COVID-19 vaccine. In this study, we aim to investigate the relationship between Twitter discourse related to the COVID-19 pandemic and the actual COVID-19 cases in different regions. This study focuses on addressing the following questions.

1. What are the prevailing sentiments expressed in COVID-19 related tweets during the pandemic?
2. What are the main topics of discussion and their frequencies in COVID -19 Twitter data?

2. Methods

This section specifically discusses the data and the analysis method to answer the research questions outlined above.

2.1 Data

The proposed study will use three datasets obtained from Kaggle [4-7]. The first dataset is the MaxMind World Cities Database, which provides a list of cities in the world [4], containing 230,000 cities, including information such as country, city, accent city, region, population, latitude, and longitude. The second dataset is the Coronavirus 2019-nCoV Data Repository by Johns Hopkins CSSE, which is a curated version of the 2019 Novel Coronavirus COVID-19 Data Repository [5], was processed by concatenating the daily reports files, adding a daily update date, and fixing country name duplicates. Additionally, missing data in latitude and longitude for Province/State and/or Country/Region was replaced using the median values of the available data for each category [6]. The dataset contains information on country/region, province/state, latitude, longitude, confirmed, recovered, deaths, and date of COVID-19 worldwide, at both country and regional levels. The third dataset is the COVID-19 Tweets dataset, was collected using Twitter API and a Python script [7]. This dataset contains tweets with the #covid19 hashtag, with collection starting on July 25, 2020, and continuing daily, including information such as username, user location, user description, user-created, user followers, user friends, user favorites, user verified, date, text, hashtags, source, is_retweet. In all, the covid data has 515800 rows and 8 columns, the tweets dataset has 179108 rows and 28 columns of which only the first 10,000 tweets will be considered, and the world cities data has 3173958 rows and 7 columns.

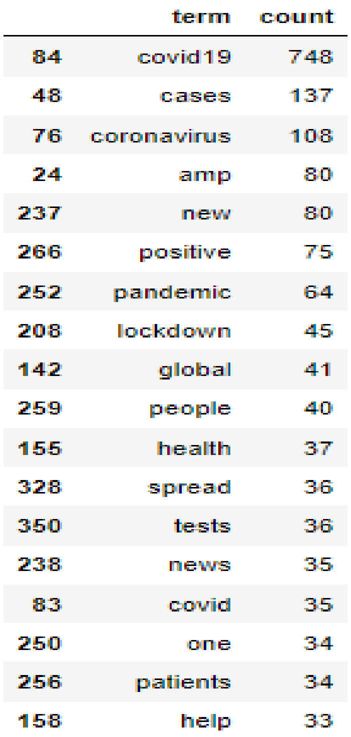
2.2 Analysis

The data will be analyzed using a combination of data manipulation, text preprocessing, and sentiment analysis techniques. The data will first be cleaned by removing URLs, usernames, hashtags,and punctuation. NLTK (Natural Language Toolkit) is a Python library used for natural language processing tasks used here for stopwords removal from the text of the tweets. The cleaned text will then be tokenized into words. Next, the text will be vectorized using the CountVectorizer class from the scikit-learn library, which will convert the text into a matrix of term frequencies. Unigrams, bigrams, and trigrams will be extracted from the text and a dataframe with the terms and their corresponding frequencies is created.

First, an instance of the Empath library was created to analyze the sentiment of each tweet across various emotions. The 27 emotions like ‘joy', 'sadness', 'anger', 'fear', 'surprise', 'love', 'trust', and 117 categories cover a wide range of topics, such as ‘health’, ‘death’, ‘violence’, ‘crime’, ‘politics’, ‘government’, ’money’, ‘business’, ‘technology’, ‘science’, ‘religion’, ‘spirituality’, ‘art’, ‘music’, ‘media’, and ‘internet’. Then, it calculates the total score for all emotions and normalizes each emotion's score by the total score. Finally, it calculates the sentiment score for each tweet by taking the average of all emotion scores. The sentiment score reflects the overall sentiment expressed in the tweet based on the analyzed emotions. A new column was added to the dataset to reflect the categories and emotions extracted from Empath. Next, a new dataframe for sentiments was created and appended to contain the text of each tweet along with its positive, negative, and neutral scores. The sentiment scores were extracted from the Empath sentiment category for each tweet by looping through the dataset and the top tweets were identified for each emotion and sentiment.

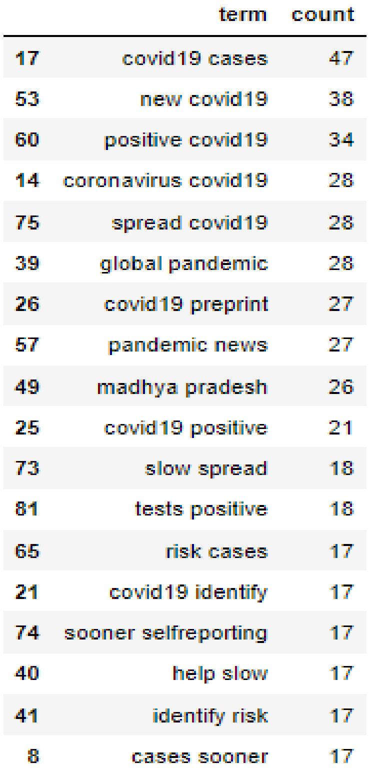
In addition, a Latent Dirichlet Allocation (LDA) model, a generative statistical model used to identify topics within a large corpus of text data, was trained using the sklearn library. The LDA model was configured to include five components and run for a maximum of 50 iterations with a random state of 0. The resulting topics and their top words were printed, and each document was assigned a topic based on its content. Through this a better understanding of the overall sentiment and emotions expressed, as well as the prevalence of different topics.

Unigrams refer to single words, so the top 5 unigrams are the 5 most common individual words like "covid19", "cases", "coronavirus", "amp", and "new". Bigrams refer to pairs of adjacent words, so the top 5 bigrams are the 5 most common two-word phrases "covid19 cases", " new covid19", " positive covid19", "coronavirus covid19", and " spread covid19".Trigrams refer to sets of three adjacent words, so the top 5 trigrams are the 5 most common three-word phrases in your dataset: "global pandemic news", " new covid19 preprint", " slow spread covid19 ", " risk cases sooner ", and " cases sooner selfreporting ". N-grams (where n refers to the number of adjacent words) are used to analyze text data and identify common patterns or topics.



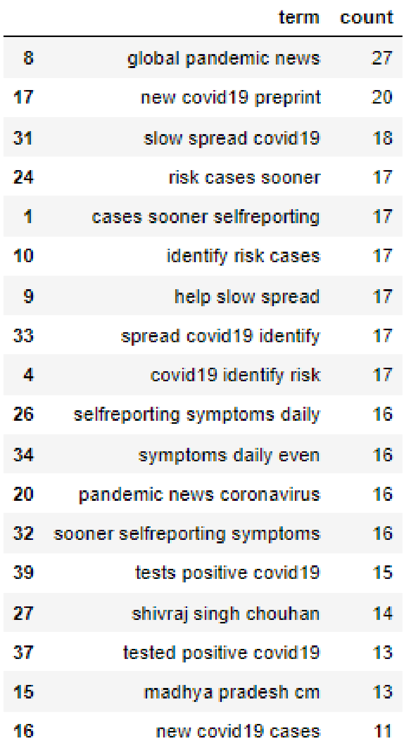
	term	count
84	covid19	748
48	cases	137
76	coronavirus	108
24	amp	80
237	new	80
266	positive	75
252	pandemic	64
208	lockdown	45
142	global	41
259	people	40
155	health	37
328	spread	36
350	tests	36
238	news	35
83	covid	35
250	one	34
256	patients	34
158	help	33

Table 1: Most popular Unigrams



	term	count
17	covid19 cases	47
53	new covid19	38
60	positive covid19	34
14	coronavirus covid19	28
75	spread covid19	28
39	global pandemic	28
26	covid19 preprint	27
57	pandemic news	27
49	madhya pradesh	26
25	covid19 positive	21
73	slow spread	18
81	tests positive	18
65	risk cases	17
21	covid19 identify	17
74	sooner selfreporting	17
40	help slow	17
41	identify risk	17
8	cases sooner	17

Table 2: Most Popular Bigrams



	term	count
8	global pandemic news	27
17	new covid19 preprint	20
31	slow spread covid19	18
24	risk cases sooner	17
1	cases sooner selfreporting	17
10	identify risk cases	17
9	help slow spread	17
33	spread covid19 identify	17
4	covid19 identify risk	17
26	selfreporting symptoms daily	16
34	symptoms daily even	16
20	pandemic news coronavirus	16
32	sooner selfreporting symptoms	16
39	tests positive covid19	15
27	shivraj singh chouhan	14
37	tested positive covid19	13
15	madhya pradesh cm	13
16	new covid19 cases	11

Table 3: Most Popular Trigrams

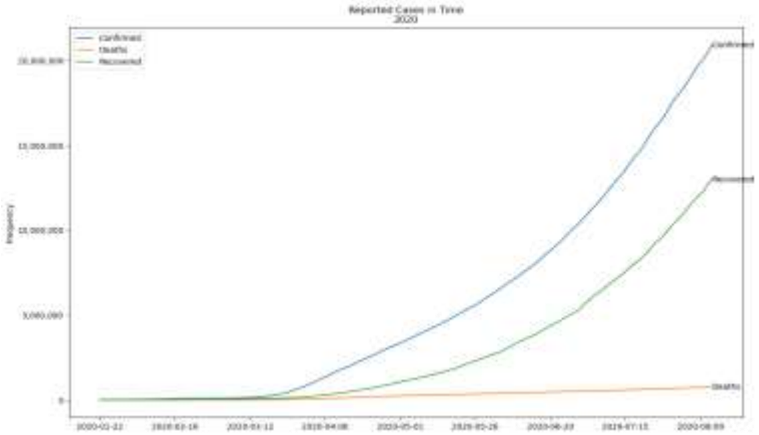


Fig 1: Reported Cases in Time – the year 2020

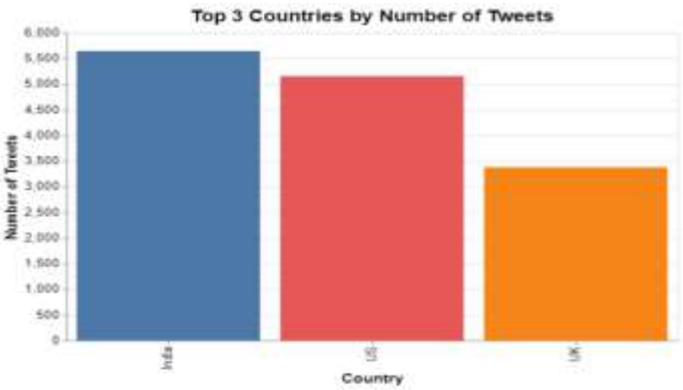


Fig 2: Top 3 countries by the number of Tweets

The top three countries mentioned based on the number of tweets, are India with 5600 tweets, the United States with 5200 tweets, and the United Kingdom with 3400 tweets.

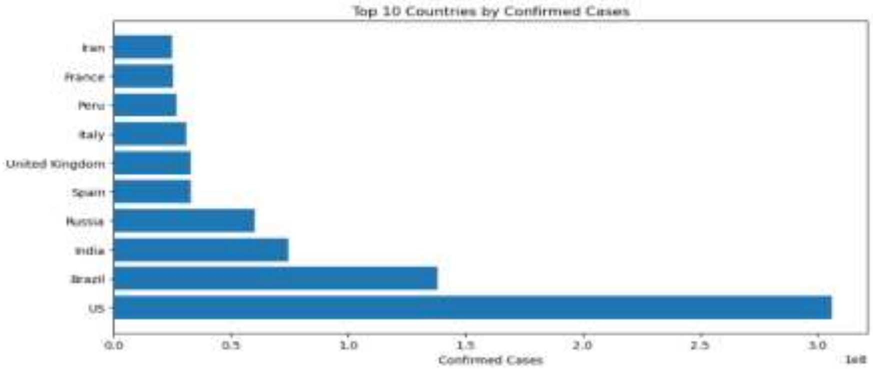


Fig 3: Top 10 countries by the number of Confirmed cases in ascending order

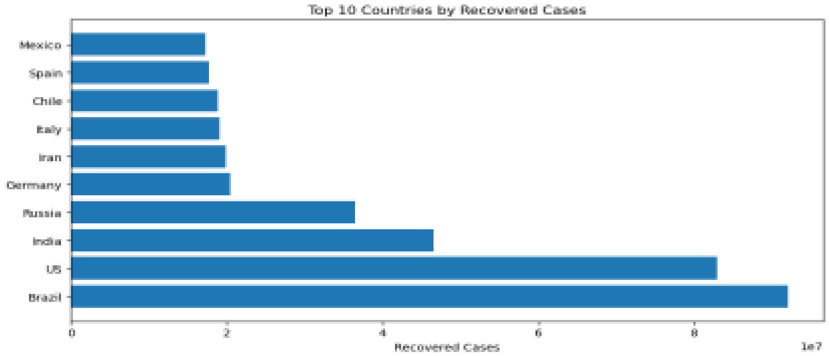


Fig 4: Top 10 countries by the number of Recovered cases in ascending order

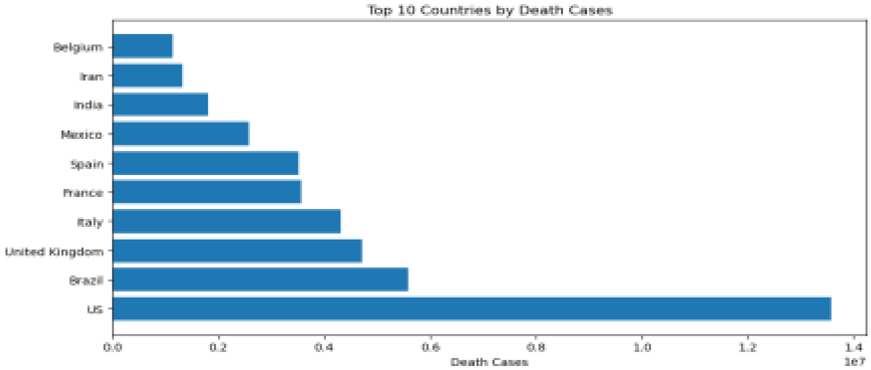


Fig 5: Top 10 countries by the number of Recovered cases in ascending order

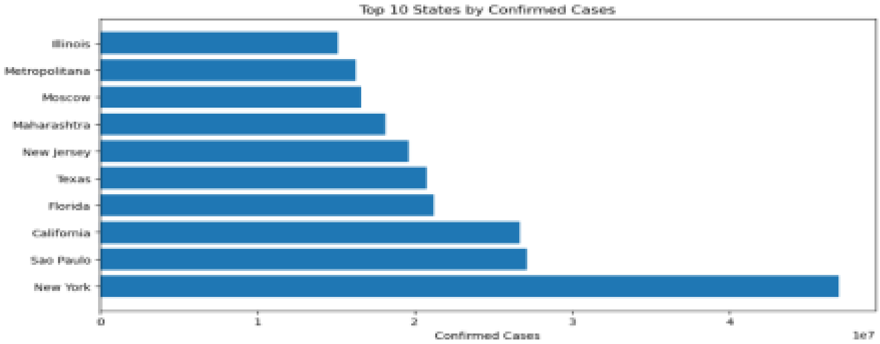


Fig 6: Top 10 States by the number of Confirmed cases in ascending order

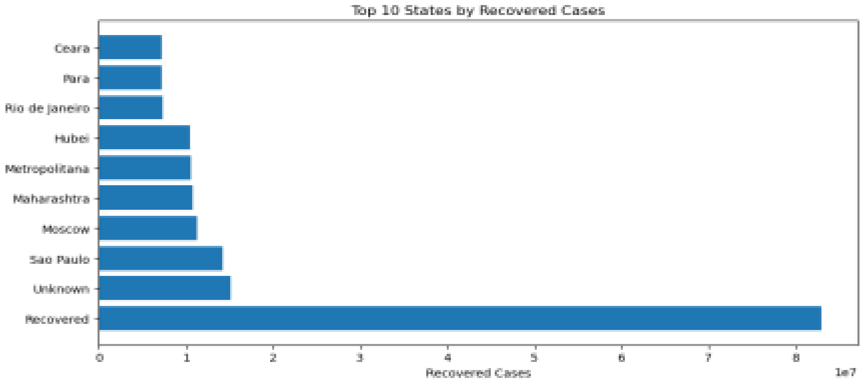


Fig 7: Top 10 States by the number of Recovered cases in ascending order

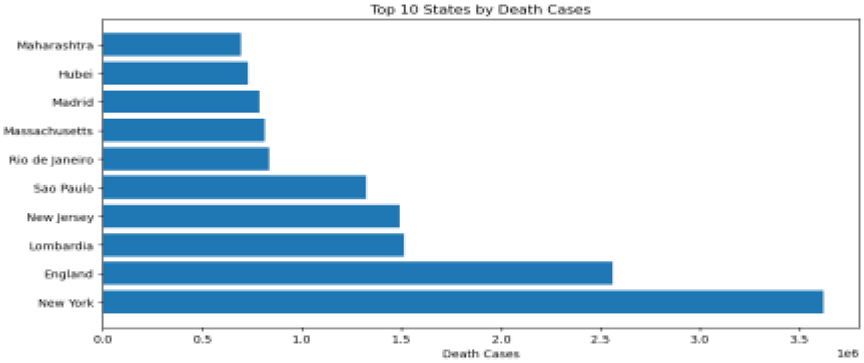


Fig 8: Top 10 States by the number of Death cases in ascending order

4 Results

Based on the sentiment analysis scores for the 10,000 tweets related to COVID-19, the most prevalent emotions expressed are sadness and fear, with scores of 369 and 333 respectively. On the other hand, joy and surprise scores are comparatively lower at 108 and 89. Love and trust scores are moderately high at 325 and 752. The anger and disgust scores are relatively low at 122 and 86 respectively, anticipation score is also moderate at 88.

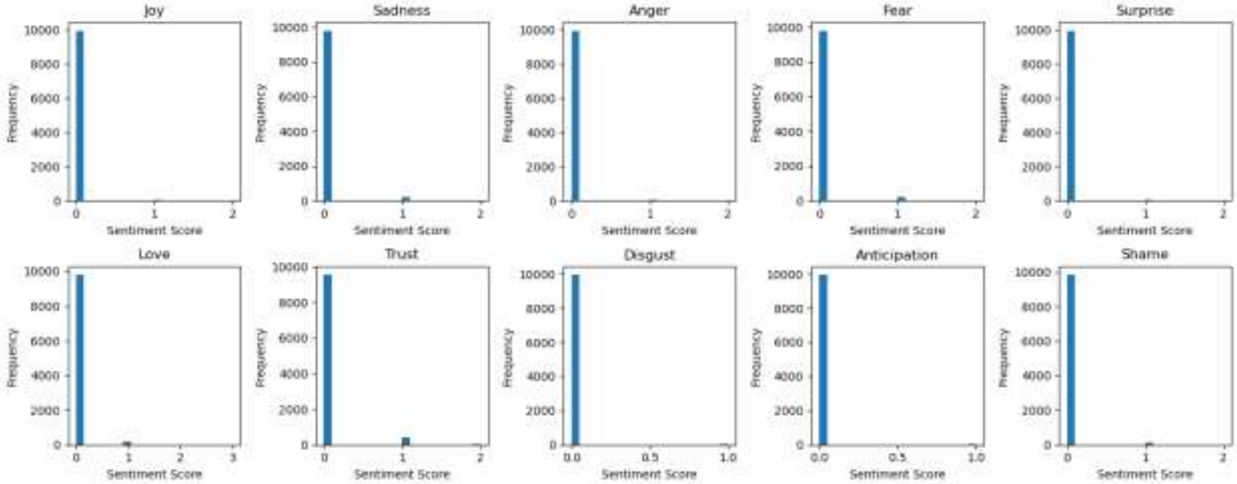


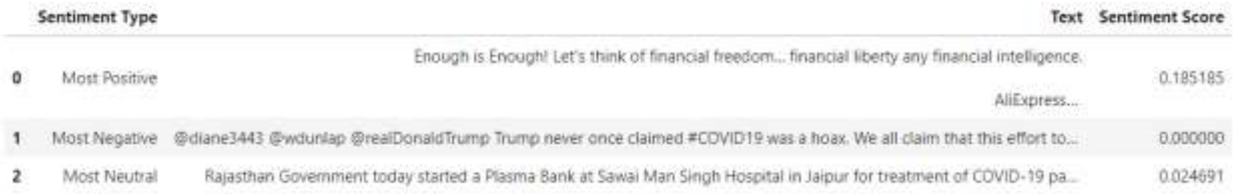
Fig 9: Bar Plots of various emotions obtained from the tweets.

	joy	sadness	anger	fear	surprise	love	trust	disgust	anticipation	shame
0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
3	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Fig 10: The top 5 tweets from the chosen emotions – joy, sadness, anger, fear, surprise, love, trust, disgust, anticipation, and shame.

	positive	negative	neutral
0	[illegible]	[illegible]	[illegible]
1	[illegible]	[illegible]	[illegible]
2	[illegible]	[illegible]	[illegible]
3	[illegible]	[illegible]	[illegible]
4	[illegible]	[illegible]	[illegible]

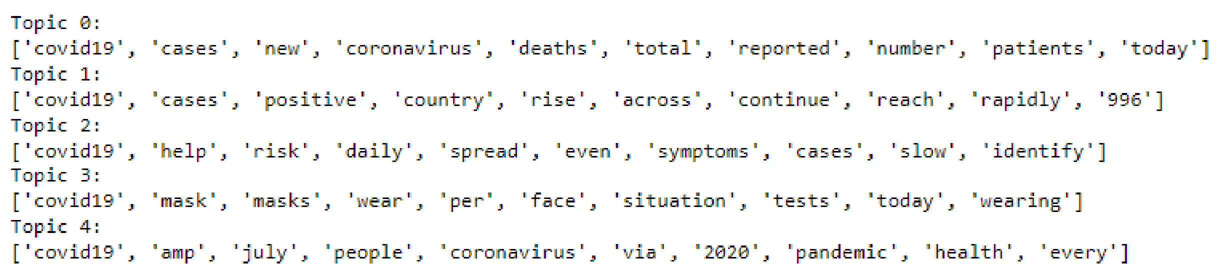
Fig 11: The top 5 tweets from the chosen sentiments – positive, negative, and neutral.



	Sentiment Type	Text	Sentiment Score
0	Most Positive	Enough is Enough! Let's think of financial freedom... financial liberty any financial intelligence. AliExpress...	0.185185
1	Most Negative	@diane3443 @wdunlap @realDonaldTrump Trump never once claimed #COVID19 was a hoax. We all claim that this effort to...	0.000000
2	Most Neutral	Rajasthan Government today started a Plasma Bank at Sawai Man Singh Hospital in Jaipur for treatment of COVID-19 pa...	0.024691

Fig 12: Tweets that are the most positive, most negative, and Neutral.

The sentiment scores of the tweets range from 0.0 to 0.185. The maximum sentiment score, which is the most positive tweet is 0.185. The minimum, also the most negative tweet is 0.0. And the average score is which is neutral is of 0.02.



```
Topic 0:
['covid19', 'cases', 'new', 'coronavirus', 'deaths', 'total', 'reported', 'number', 'patients', 'today']
Topic 1:
['covid19', 'cases', 'positive', 'country', 'rise', 'across', 'continue', 'reach', 'rapidly', '996']
Topic 2:
['covid19', 'help', 'risk', 'daily', 'spread', 'even', 'symptoms', 'cases', 'slow', 'identify']
Topic 3:
['covid19', 'mask', 'masks', 'wear', 'per', 'face', 'situation', 'tests', 'today', 'wearing']
Topic 4:
['covid19', 'amp', 'july', 'people', 'coronavirus', 'via', '2020', 'pandemic', 'health', 'every']
```

Fig 14: The top 5 topics of COVID -19 Twitter data obtained from topic modeling using LDA.

The results show the top words in each of the five topics that were identified through the Latent Dirichlet Allocation (LDA) algorithm applied to the preprocessed COVID-19 related tweets. Topic 0, with a frequency of 4010, appears to be about the statistics related to the COVID-19 pandemic, including the number of cases, deaths, and patients reported. Topic 1, with a frequency of 2203 seems to be about the rapid rise in COVID-19 cases across different countries. Topic 2 with a frequency of 12998 is about the measures taken to slow the spread of COVID-19, including identifying symptoms and reducing risks. Topic 3 with a frequency of 703 is about the importance of wearing masks in the current COVID-19 situation and the need for testing. Topic 4 with a frequency of 6073, seems to be a more general topic that talks about the COVID-19 pandemic and its impact on health and people in general.

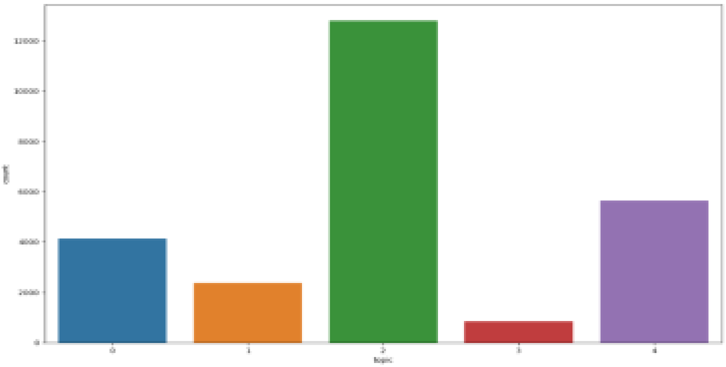


Fig 14: The frequency of the top 5 topics of COVID -19 Twitter Data.

5. Conclusion and Limitations

The prevailing sentiments like sadness and fear, indicate that people are experiencing negative emotions due to the pandemic's impact on their lives. Love and trust scores were moderately high, suggesting that people have trust and affection toward healthcare workers and government officials working to combat the pandemic. The anger and disgust scores are relatively low, implying that people are not expressing these emotions as frequently or they are expressing them in a more restrained manner. The anticipation score is also moderate, indicating people are hopeful or anxious about the future course of the pandemic. An average score is 0.02 means that the overall sentiment of the tweets is slightly positive but mostly neutral. It is also possible that people are trying to maintain a positive outlook despite the challenges posed by the pandemic. The results of the study also identified the main topics of discussion in COVID-19 Twitter data, including statistics related to the pandemic, the rapid rise in cases, measures taken to slow the spread of COVID-19, the importance of wearing masks and the pandemic's impact on health and people in general.

The study has limitations and potential sources of bias, such as the use of a lexicon-based approach like empathy relies on pre-defined categories that may not capture human language nuances and context.

Additionally, the validity of the study depends on the representativeness of the sample of tweets and the generalizability of the findings to other populations. Further, analyzing COVID-19 tweets may be affected by the informational nature of tweets, tweets in multiple languages or regions, and the limitations of the sentiment analysis algorithm like Empath.

6 References

[1] Hussain, A., Ali, S., Ahmed, M., & Hussain, S. (2020). The outbreak of Coronavirus Disease (COVID-19) in China: A systematic review and meta-analysis. Journal of Infection and Public Health, 13(5), 742-748. doi:10.1016/j.jiph.2020.03.019

[2] Li, C., Chen, L., Chen, X., Zhang, M., & Pang, J. (2020). Using Twitter data to predict COVID-19 cases at the county level in the United States. PLoS ONE, 15(10), e0240074. doi:10.1371/journal.pone.0240074

[3] Singh, L., Bansal, S., Bode, L., Budak, C., Chi, G., Kawintiranon, K., . . . Vraga, E. (2021). A first look at COVID-19 information and misinformation sharing on Twitter: Results from the first seven weeks. Harvard Kennedy School Misinformation Review, 1(1). doi:10.37016/mr-2020-048

[4] M. Mind, “World Cities Database,” Kaggle, 23-Aug-2017. [Online]. Available: https://www.kaggle.com/datasets/max-mind/world-cities-database. [Accessed: 08-Apr-2023].

[5] C. S. S. E. at J. H. U. CSSEGISandData, “Covid19/csse_covid_19_data/csse_covid_19_daily_reports at master · CSSEGISANDDATA/covid-19,” GitHub, 2019. [Online]. Available: https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_daily_reports. [Accessed: 23-Apr-2023].

[6] A. N. D. R. A. D. A. Andradaolteanu, “🦠COVID-19: Sentiment analysis &amp; Social Networks,” Kaggle, 23-Sep-2020. [Online]. Available: https://www.kaggle.com/code/andradaolteanu/covid-19-sentiment-analysis-social-networks/input. [Accessed: 23-Apr-2023]..

[7] G. Preda, “Covid19 tweets,” Kaggle, 30-Aug-2020. [Online]. Available: https://www.kaggle.com/datasets/gpreda/covid19-tweets. [Accessed: 08-Apr-2023].