Effects of Covid-19 on Mental Health

Ritvik R. Prabhu[a)]
Data Bridge, Virginia Tech, Blacksburg, Virginia, USA

(Dated: 27 July 2021)

In late 2019, the Chinese government reported an outbreak of a pneumonia-like virus in Wuhan, China. This virus was named SARS-cov-2 or Covid 19. Ever since, the number of cases where people have contracted the virus has exponentially grown, to the point where nations implemented lock-downs to control the spread of the virus. This has led to a drastic change in lifestyle where socialization was brought to a minimum, people's jobs are at stake, small businesses' success is threatened, and the world has moved almost entirely online. Sudden changes in lifestyle with no evident end in sight could have effects on mental health. The hypothesis for the study is that the public's general mental health gets worse as the number of cases rise, and there are restrictions placed that drastically change the lifestyle of the people. This study is meant to help visualize the impact of this new lifestyle on mental health over time and correlate this data with the significant events that have taken place around the same timeline. To achieve this, the Reddit API was used to scrape for information through subreddits that are popular mental health forums. Showing the change in the subreddit activity was impossible due to limitations in the API. The data gained was biased, and hence no conclusive results were found. The next part of the experiment involved scraping data from popular covid subreddits that do not bias mental health. Sentiment analysis was performed on the data. Since the data was predominantly mentions of scientific study rather than a discussion of mental health, the results were also non-conclusive. Furthermore, it was observed that NLP tools might have to be tweaked to perform a sentiment analysis on medical terminology as testing positive for the virus is considered a positive reading.

Keywords: Covid-19, Sentiment Analysis, Mental Health, Reddit, PRAW, NLTK

I. INTRODUCTION

The novel COVID-19 outbreak in Wuhan, China has led to a significant changes on our lifestyle. These changes have a significant impact on the mental health, as many people experience loss of income, social interaction, mobility and physical health. The uncertainty of the duration of this new lifestyle, adds on to the strain on mental health. It is necessary to develop a clear model that helps visualize the impact on mental health in order to develop appropriate interventions and solution for the general population and at-risk groups[1].

Social media is a great place to use as a data set to analyze the status of mental health of the general public. Because of dissociative Anonymity, people tend to be more open online since there is no need to take any ownership of your actions online[2]. Public commentary posted on mental health support groups on the popular social media website- Reddit- can be used for this very purpose.

Reddit is a massive collection of forums where people can share news and content related to the forums. Reddit is broken up into more than a million communities known as "subreddits," each of which covers a different topic. Each subreddit begins with the prefix 'r/'[3]. For this study we will be focused on those subreddits that are focused on mental health.

The programming language chosen for this research was Python. The language was chosen under the conditions that there is a low learning curve and is a very powerful object oriented programming language.

Natural Language Processing (NLP) is manipulation or understanding text or speech by any software or machine. This is the concept used during sentiment analysis. A powerful tool used for NLP is Python's NLTK library. This toolkit is one of the most powerful NLP libraries which contains packages to make machines understand human language and reply to it with an appropriate response. Tokenization, Stemming, Lemmatization, Punctuation, Character count and word count are just some of the functions on this library[4].

Using NLP on social media posts to analyze the sentiment is a growing field[5,6]. On Reddit, users post anonymous posts on subreddits. This data is immediately publicly available. This data allows for comparisons of multiple time frames and creates a documentation of first person experiences[7]

In this study, we monitor several popular subreddits that are online mental health support groups, and apply text-mining and sentiment analysis to understand the state of mental health ever since the beginning of quarantine (in the US). According to a study undertaken by MIT, there was a rise in social media posts regarding Covid 19 during March, 2020[7]. The hypothesis for the study is that the public's general mental health gets worse as the number of cases rise, and there are restrictions placed that drastically change the lifestyle of the people.

[a)]Electronic mail: ritvikp@vt.edu

II. METHODS

A. Independent Subreddit Analysis

This section narrates the analysis techniques used for the individual subreddits focused around mental health.

1. *Data Downloading and processing*

Using the PRAW API, data from the following subreddits were downloaded:-

- r/addiction
- r/anxiety
- r/depression
- r/EDAnonymous
- r/healthanxiety
- r/lonely
- r/ptsd
- r/spcialanxiety
- r/sucidewatch

The data was then filtered using the search queries:

- COVID-19
- Coronavirus
- Pandemic
- Quarantine
- isolation
- Social distancing
- Rona
- Covid
- 2020
- SARS-cov-2
- disease
- virus
- sars
- masks
- vaccines

The results after filtration showed posts that contain Covid related keywords. The keywords above is supposed to cover a variety of genres under the broad canopy of COVID-19 in order to broaden the dataset. (for example, it covers genres of academia by using keywords such as "SARS-cov-2" and even social media lingo like "Rona"). The data is sorted from newest to oldest and has a time filter 'all'[8].

The data was then formatted into a pandas data frame for ease in evaluation.

2. *Data Analysis*

Using the data that was extracted, we created a time series model that depicts the total number of posts with Covid related keywords from all of the subreddits against the month of the year.

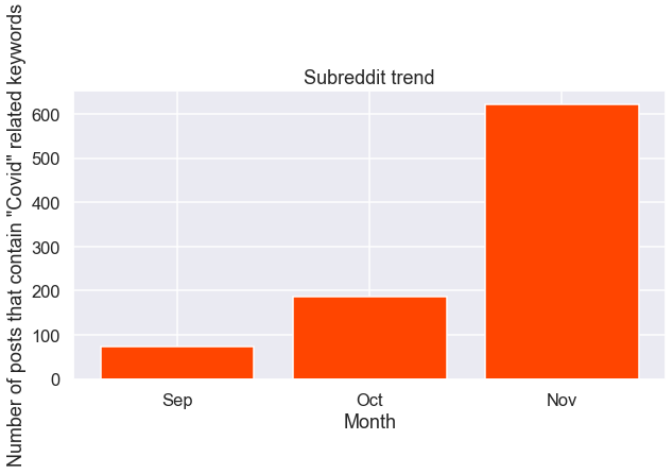

FIG. 1. Graphical representation of the total number of posts with Covid related keywords from all of the subreddits against the month of the year

As we can see in Figure 1, the results from parsing through specific subreddits led to inconclusive results, we are going to parse through the most popular subreddit for the Coronavirus - "COVID19" - in order to assess the positive, negative and neutral sentiments of the subreddit.

B. r/COVID19 Subreddit Analysis

Since the previous analysis led to inconclusive results, we decided to take it a step further and perform a sentiment analysis on a popular subreddit for Covid-19, to understand the general sentiment within the subreddit and also tokenize words to give evidence for the aforementioned analysis.

1. *Data Downloading and processing*

We parse through specifically downloaded data from r/COVID19. The data from r/COVID19 is pulled and this data is iterated through to obtain the headlines of each post pulled.

2. *NLP and Sentiment Analysis*

The library used for NLP and Sentiment Analysis is NLTK. Vader lexicon and stop words are downloaded from NLTK. Vader lexicon is a sentiment analysis tool that specifically attuned to sentiments expressed in social media[9].

Using sentiment analyzer we labeled the data as positive, negative and neutral. We iterated through the lines and use the polarity scores method to get the sentiment score. This data is ultimately added to a data frame. We then labeled the data with compound value greater than 0.1 as positive and less than -0.1 as negative.

The following is an example of the positive, neutral and negative results:-

- *Positive headline: The UK RECOVERY trial: An ambitious test of COVID-19 treatments.*
- *Neutral headline: Healthcare workers 7 times as likely to have severe COVID-19 as other workers.*
- *Negative headline: Anti-SARS-CoV-2 activity of Andrographis paniculata extract and its major component.*

As we can see, the model is not entirely accurate.

3. *Statistical Analysis*

Using the data, we performed some statistical analysis. We plot a graphical representation of the percentage values of the positive, negative and neutral posts.

As you can see in Figure 2, the results here are not only biased but also non-conclusive since there is a flaw in the NLP tool and the positive and negative percentage values are very close.

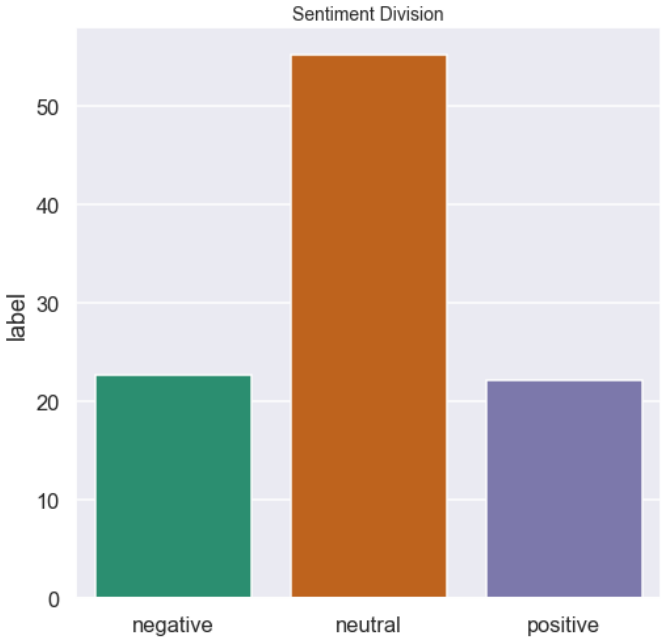

FIG. 2. Graphical representation of the percentage values of the positive, negative and neutral posts

4. *Tokenizer*

Here we analyzed the most common words in r/COVID19 using tokenizer. To begin, we have to normalize the data by making sure that the data is free from any stop words and punctuation.

The following are the words and their respective frequencies obtained from the analysis:

- covid: 262
- 19: 252
- 2: 200
- sars: 198
- cov: 192
- patients: 57
- infection: 48
- study: 40
- vaccine: 39
- coronavirus: 33

As we can see, this data indicates that the sentiment leans towards it being neutral. This is reflected by the sentiment analyzer. While the sentiment analyzer is slightly accurate, there are clear examples shown above that prove otherwise.

III. CONCLUSION

On performing an analysis on select subreddits, the results were inconclusive because of the fact that the API limits the access to 1000 posts at a time. Hence, data prior to November is not accessible and hence not available to make an accurate analysis.

On performing a sentiment analysis on r/COVID19 it was evident that the data was primarily news articles of political and scientific breakthroughs rather than a forum for discussion of mental health. This along with the fact that the NLP tool is not suited for medical terminology, the analysis led to inconclusive results because of several inaccuracies and lack of data.

In conclusion, because of the bias in data, lack of access to data and the limitations of the NLP tool performing a sentiment analysis on medical terminology, the results of this research is inconclusive.

IV. LIMITATIONS

The following are limitations that were run into over the course of this research:-

- The search terms may be stemmed. A search for "dogs" may return results with the word "dog" in them.
- Search results are limited to 1000 results. This limitation should change soon since Reddit is planning on allowing data dump through offline systems, as online systems get pretty backed up with traffic[10]
- The sentiment analyzer is not very accurate. In certain cases, it would even consider testing positive for the virus as a good thing, when in reality it is not.
- Some of the data available could be a result of self diagnosis which could taint the integrity of the data.

V. FUTURE WORK

By vectorizing the sentiment data, we can superimpose this data on a time series graph along with current events to map out key possibilities for the behaviour of the data. We can then use Machine Learning models (like Long Short Term Memory) to predict the aggregate mental health by using current events and also historical as parameters for the output. This ML model would constantly train itself of recent data in order to increase its accuracy.

ACKNOWLEDGMENTS

[1] V. Giallonardo, G. Sampogna, V. Del Vecchio, M. Luciano, U. Albert, C. Carmassi, G. Carrà, F. Cirulli, B. Dell'Osso, M. G. Nanni, and et al., "The impact of quarantine and physical distancing following covid-19 on mental health: Study protocol of a multicentric italian population trial," Frontiers in psychiatry (2020).

[2] J. Suler, "The online disinhibition effect," Cyberpsychology behavior : the impact of the Internet, multimedia and virtual reality on behavior and society **7**, 321–6 (2004).

[3] J. Widman, "What is reddit?" (2020).

[4] "Nltk (natural language toolkit) tutorial in python,".

[5] S. C. Guntuku, D. B. Yaden, M. L. Kern, L. H. Ungar, and J. C. Eichstaedt, "Detecting depression and mental illness on social media: an integrative review," Current Opinion in Behavioral Sciences **18**, 43 – 49 (2017), big data in the behavioural sciences.

[6] R. A. CALVO, D. N. MILNE, M. S. HUSSAIN, and H. CHRISTENSEN, "Natural language processing in mental health applications using non-clinical texts," Natural Language Engineering **23**, 649–685 (2017).

[7] D. M. Low, L. Rumker, T. Talkar, J. Torous, G. Cecchi, and S. S. Ghosh, Natural Language Processing Reveals Vulnerable Mental Health Support Groups and Heightened Health Anxiety on Reddit During COVID-19: Observational Study (2020).

[8] "reddit.com," ().

[9] N. Data, "Vader lexicon," (2017).

[10] "r/ideasfortheadmins - comment by u/spladug on "ever wondered the data liberation policy of reddit?"," ().