Twitter Data Sentiment Analysis to Understand the Effects of COVID-19 on Mental Health

Adeola Adetokunbo Ayandeyi* and Baidya Nath Saha
Department of Mathematical and Physical Sciences, Concordia University of Edmonton, Alberta, T5B 4E4, Canada

* Corresponding author

doi:https://doi.org/10.21467/proceedings.115.23

ABSTRACT

Coronavirus pandemic has caused major change in peoples' personal and social lives. The psychological effects have been substantial because it has affected the ways people live, work, and even socialize. It has also become major discussions on social media platforms as people showcase their opinions and the effect of the virus on their mental health particularly. This pandemic is the first of its kind as humans has never encountered anything like this virus. Handling it was very difficult at first as its characteristics are peculiar. Eventually, it was detected that it is airborne and so there is need to social distance. Before the virus surfaced, some countries of the world were dealing with mental health cases, with over 40 percent of adults in the USA reported experiencing mental health challenges, including anxiety and depression. Social media has become one of the major sources of information due to information sharing on a very large scale. People perception and emotions are also portrayed through their conversations. In this research work, the interaction and conversation of people on social media, particularly Twitter, will be analyzed using machine learning tools and algorithm to determine the effect of the virus on the mental health of people and help suggest the area of concentration to medical practitioners in order to speed up the recovery process and reduce the mental health issues which has escalated due to the virus.

Keywords: Coronavirus, pandemic, machine learning, mental health, virus.

I. Introduction

Coronavirus took the world by surprise as the first case with noted in November, 2019 in Wuhan, China and has spread to almost all the countries of the world. Since the appearance of this virus, with the source still unknown, it has been known to be air borne and the symptoms does not materialize until after few days. This means a carrier can infect people without knowing that he or she has the virus. Before this was detected, the virus had spread to many people and other countries. Social distancing was one way to reduce the spread. The increase in the number of confirmed cases and the quick spread of the virus led to about 947500 deaths globally and about 30,380,035 cases globally (as at 18th September, 2020) [1]. Although the increase in confirmed cases and deaths are no longer at its peak, its effect on human life and activities has taken a drastic turn. Social distancing has been inculcated into daily activities, unemployment increased drastically, school closure is still very paramount in most nations and the economy of most nations are suffering. All these have in more ways than one, affected the mental health of human beings. Humans are social beings, the effect of this virus has led to loneliness, depression and increase in mental health cases.The recent survey carried out by Census Bureau and the Centers for Disease Control and Prevention shows that coronavirus is associated with rapid rises in psychological distress across many nations most especially among women, the less educated and some minority ethic groups like black Americans. The fear of the virus has also triggered new mental illnesses which means that the practical impact of the crisis is way more than the actual number of infection cases or fatalities [14].A lot of conversations and emotional expressions on the virus

Proceedings DOI: 10.21467/proceedings.115; Series: AIJR Proceedings; ISSN: 2582-3922; ISBN: 978-81-947843-5-7

and its effect on peoples' lives has becoming a very popular topic on social media. In this research work, we will be analyzing these conversions and expressions in order to determine the areas where people are most affected by coronavirus and help the medical practitioners and the government narrow down to the specific areas to look into in order to reduce the effect of the virus on peoples' lives.

II. Literature Review

There are quite a number of research work that focus on social media analysis in order to depict people opinions and perspective and also to gather relevant data. Social media has a load of information shared which can be scraped and analyzed.HJ Do,et al investigated people's emotional responses expressed on Twitter during the 2015 Middle East Respiratory Syndrome (MERS) outbreak in South Korea [6]. Mansur ALP et al conducted emotional analysis of Turkish tweet using Deep Neural Networks and classified the data into six basic emotions which are joy, sadness, anger, fear, disgust, and surprise [7]. H. Achreka predicted flu trends using Twitter data. Felipe Taliar Giuntini et al Identified Emotional Expressions on Facebook Reactions Using Clustering Mechanism [8]. Man Hung et al conducted a research where social media discussions related to COVID-19 were analysed to investigate social sentiments toward COVID-19–related themes [11]. The goal of the study was to provide clarity about online COVID-19–related discussion themes and to examine sentiments associated with COVID-19 [11]. Koustuv Saha et al carried out a research on the psychosocial effects of the COVID-19 crisis by using social media data (Twitter) from 2020. In their research, they found out that people's mental health symptomatic and support expressions increased significantly during the COVID-19 period as compared to similar data from 2019 [13]

III. Methodology

For this research work, Data collection method and analysis is used. Covid 19 and mental health related data is retrieved from twitter related to Covid-19, the data is structured, cleaned to removed unwanted data and Natural Language Processing module(nltk) for text processing is applied on the structured data in order to determine the word frequencies and also categorize the text into positive, negative and neutral sentiments using nltk module called Vader. All these are carried out with python programming language.

IV. Experimental Results and Discussions

A. Data Collection

Data is collected using the twarc module of python. The unique ids of twitter users with conversations about covid-19 and mental health are retrieved from Zenodo database [15]. All the tweet information for each ids such as the text, date, location, language and other information were extracted from twitter starting from January till October 2020 using the twarc module of python programming language. Due to the size of the data collected, a sample of about 4 million rows of data was used for this research work. These data are stored in a dataframe and saved as a csv file. The columns of the dataframe are text, id, and the time. The 'text' column stores the information that users provide about the subject in question. These are user's opinion, fact, information and discussion about covid-19 and mental health. The 'id' column stores the unique identity of twitter user whose text are analyzed. The 'time' column stores the exact time tweets were posted on twitter by the user. This helps to analyze the number of tweets at a given time and at what time of the day tweets were at the peak about COVID-19 and mental health.

B. Data Analysis

Data retrieved from twitter requires cleaning and this is done by removing stop words, special characters and words that does not add any meaning to the conversations. With the cleaned data of 1409754 tweets and 63,424,231 words, the frequency of tweets per month and a wordcloud is shown in Fig.1 and Fig.2 respectively to visualize the frequencies of keywords.

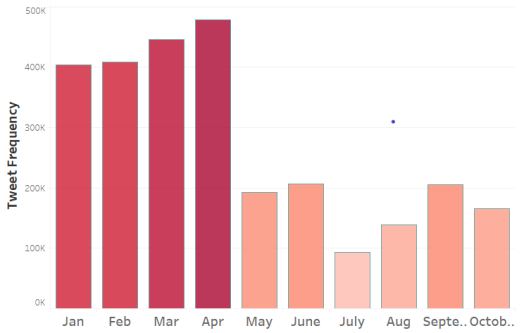

Fig. 1: Monthly Tweet Frequency

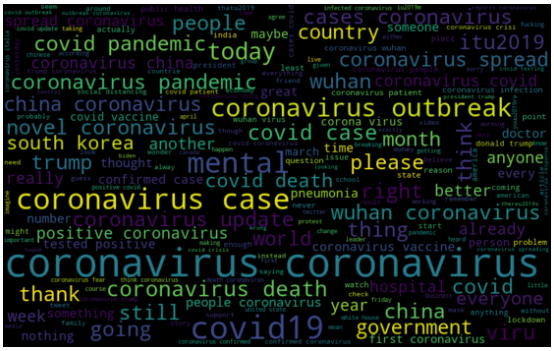

Fig. 2: Keyword with most frequency

Sentimental analysis was conducted on the cleaned text (extracted keywords) to retrieve and compare with list of words associated with certain emotions [5], SentimentIntensityAnalyzer module of nltk of Natural Language Processing in python is used for this purpose. The frequency of these sentiments is plotted as a bar chart shown in Fig.3 to visualize these sentiments and people's perspective about covid-19 and its effect on mental health based on their conversations in their tweets.

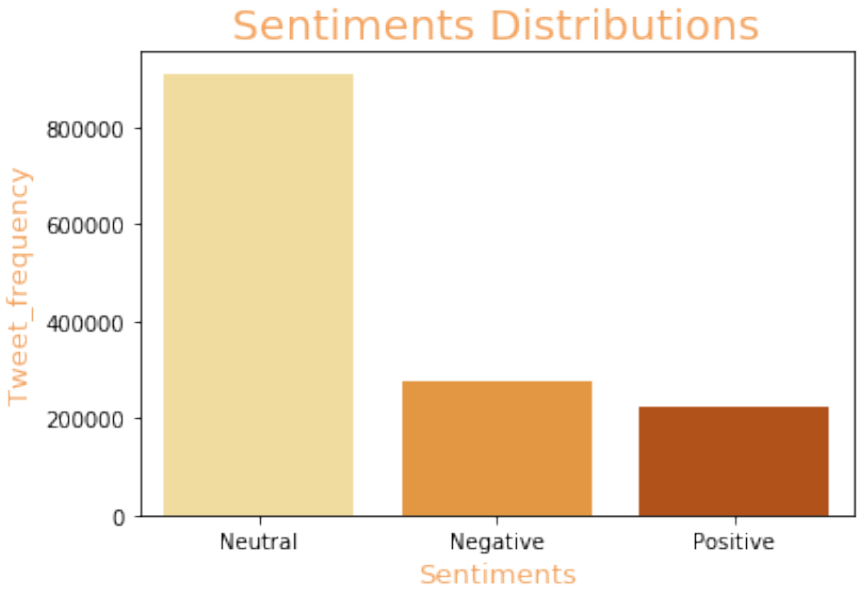

Fig. 3: Sentiment Distribution

The positive and negative sentiments were extracted from the tweets frequency and wordclouds were generated to visualize the words with positive sentiments and with negative sentiments. Fig.4 and Fig.5 displays positive and the negative wordclouds respectively.

Fig. 4: Positive Sentiments

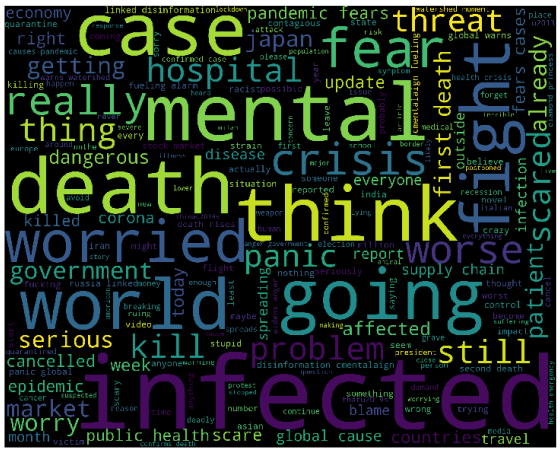

Fig. 5: Negative Sentiments

To further understand the perception of people based on their tweets, semi-supervised learning algorithm [12] was utilized to create different models that can further assist with the analysis of peoples' expression. Six modules of machine learning was used to create the model and each was compared to determine the model with the least Root Mean Square Error RMSE). These algorithms are RandomForestRegressor, XGBRegressor, Ridge, BayesianRidge, ExtraTreesRegressor, ElasticNet, KNeighborsRegressor and GradientBoostingRegressor. The model with the least value of RMSE is said to be the most appropriate model to use. Fig 6 shows the result of RMSE for each model

```
RandomForestRegressor      CV-5 RMSE: 0.62 (+/- 0.00)
XGBRegressor               CV-5 RMSE: 0.62 (+/- 0.00)
Ridge                      CV-5 RMSE: 0.78 (+/- 0.00)
BayesianRidge              CV-5 RMSE: 0.78 (+/- 0.00)
ExtraTreesRegressor        CV-5 RMSE: 0.61 (+/- 0.00)
ElasticNet                 CV-5 RMSE: 0.78 (+/- 0.00)
KNeighborsRegressor        CV-5 RMSE: 0.70 (+/- 0.00)
GradientBoostingRegressor CV-5 RMSE: 0.69 (+/- 0.00)
```

Fig. 6: RMSE for different models

Further analysis was carried out using PseudoLabeler algorithm to create an additional model. This further determined the model with the lower RMSE values and also created RMSE values for a sample data. As the sample rate increases, the RMSE values for RandomForestRegressor decreases while the RMSE values for XGBRegressor increases. This shows that the most appropriate model to use is RandomForestRegressor. Fig.7 below shows the new RMSE values for RandomForestRegressor and XGBRegressors after going through the PseudoLabeler algorithms

```
XGBRegressor        CV-8 MSE: 0.5207 (+/- 0.0094)
PseudoLabeler       CV-8 MSE: 0.5234 (+/- 0.0082)
RandomForestRegressor
XGBRegressor {'RandomForestRegressor': [0.5273477648162085, 0.5248872771253827,
0.5257213829420355, 0.5247266873524229, 0.5247275178388064, 0.5243937488088508,
0.5250399013295511, 0.5236977298150265, 0.5229540941666622, 0.5226495338957032],
'XGBRegressor': [0.5235778390100007, 0.5232973514503036, 0.5249746922415829,
0.5265316692048296, 0.5265803886265391, 0.5279308806858164, 0.5267381964481872,
0.5280905923625739, 0.5271881943225006, 0.5291283502358649]}
```

Fig. 7: Better RMSE for RandomForestRegressor and XGBRegressor

From the results of the above models, there are quite some results which shows the rate at which the RMSE increases or decreases as the sample rate increases. RandomForestRegressor had a better result because the RMSE reduces as the sample rate increases. Fig.8 depicts this result.

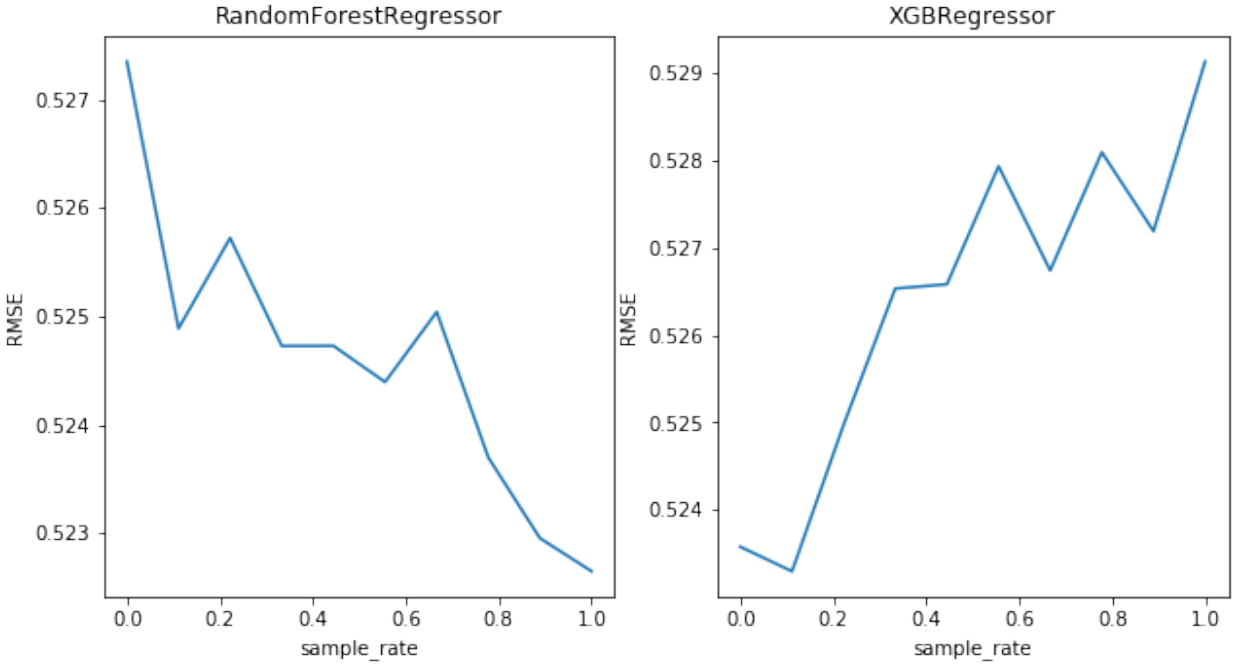

Fig. 8: Better RMSE for RandomForestRegressor and XGBRegressor

V. Conclusion

From the study and analysis of twitter data using Natural Language Processing and Scikit machine learning algorithm to create AI models, we deduced that the mental health issues increased between January and April but gradually reduced as people tends to manage the situation. The trend of Covid 19 and its effect on mental health kept fluctuating as the months go by but still within minimized range. The effect of this pandemic on mental health is more centered around personal challenges such as unemployment, isolation from loved ones, travel restrictions and other unforeseen circumstances. This study will assist medical professionals to focus on the specific challenges that triggers the mental issues in patients and help with resolutions as quickly as possible. The research is still in proress as we keep analysing the effect of this virus on mental health using social media data so as to bring about last solution to mental health issues.

Proceedings DOI: 10.21467/proceedings.115
ISBN: 978-81-947843-5-7

Series: AIJR Proceedings
ISSN: 2582-3922

REFERENCES

[1] "Coronavirus Cases:" Worldometer. [Online]. Available: https://www.worldometers.info/coronavirus/

[2] Preeti Malani, Chief Health officer at University of Michigan Medicine

[3] Arianna Huffington, founder of Huffington Post and CEO of Thrive Global

[4] Md. Yasin Kabir, Sanjay Madria "CoronaVis: A Real-time COVID-19 Tweets Data Analyzer and Data Repository"

[5] Amrita Mathur, Purnima Kubde "Emotional Analysis using Twitter Data during Pandemic Situation: COVID-19"

[6] H. J. Do, C. G. Lim, Y. J. Kim and H. J. Choi, "Analyzing emotions in twitter during a crisis: A case study of the 2015 Middle East Respiratory Syndrome outbreak in Korea", 2016 Int. Conf. Big Data Smart Comput. BigComp 2016, pp. 415-418, 2016.

[7] M. A. Tocoglu, O. Ozturkmenoglu and A. Alpkocak, "Emotion Analysis from Turkish Tweets Using Deep Neural Networks"

[8] H. Achrekar, A. Gandhe, R Lazarus, S. H. Yu and B. Liu, "Predicting flu trends using twitter data", 2011 IEEE Conf. Comput. Commun. Work INFOCOM WKSHPS 2011, pp. 702-707, 2011.

[9] F. T. Giuntini et al., "How do i feel? Identifying emotional expressions on Facebook reactions using clustering mechanism", IEEE Access, vol. 7, pp. 53909-53921, 2019.

[10] Sai Teja, "Stop Words in NLP".[Online]. Available: https://medium.com/@saitejaponugoti/stop-words-in-nlp-5b248dadad47.

[11] Man Humg et al, 'Social Network Analysis of COVID-19 Sentiments: Application of Artificial Intelligence'

[12] "Pseudo Labeling: Semi Supervised Learning," Analytics Vidhya, 07-Jun-2020. [Online]. Available: https://www.analyticsvidhya.com/blog/2017/09/pseudo-labelling-semi-supervised-learning-technique/.

[13] Koustuv Saha et al "Social Media Reveals Psychosocial Effects of the COVID-19 Pandemic"

[14] McGinty EE, Presskreischer R, Han H, Barry CL. Psychological Distress and Loneliness Reported by US Adults in 2018 and April 2020. JAMA [Internet] 2020

[15] "Research. Shared.," Zenodo. [Online]. Available: https://zenodo.org/record.