Exploring the Effect of the COVID-19 Pandemic on Mental Health through Natural Language Processing of Tweets

Alden Felix

University of Texas at Dallas

Abstract

The onset of the COVID-19 Pandemic was marked by sudden and drastic changes to daily life. These changes combined with the worsening global health crisis severely impacted the mental health of many people. This study seeks to understand the ways in which people's mental health was impacted while also evaluating the effectiveness of Twitter (X) tweets as an alternative method of traditional public survey. To this end I employ natural language processing techniques consisting of topic modeling and a dictionary method on a sample of tweets taken from the Dallas Forth Worth Metroplex during 3 periods: before the pandemic, the start of the pandemic, and a year after the start of the pandemic. Although I was not able to conclude the effect of the pandemic on mental health from this study, my research provides insights on the effective use of tweets in natural language processing for future research to build upon.

Keywords

COVID-19, pandemic, mental health, machine learning, natural language processing, topic modeling, dictionary method

Introduction

On December 1st, 2019, the first COVID-19 case is detected in Wuhan, China. On January 30th, 2020, the World Health Organization (WHO) declares COVID-19 as an international health crisis. On March 11th, 2020, three months after the initial detection of COVID-19, WHO declares a pandemic. Governments and organizations around the world implement drastic measures to combat the spread, foremost of which are nationwide social isolation and distancing measures. (Siddiqui et al., 2022). The sudden and drastic changes led to conditions that worsened the mental health of many individuals. Surveys taken to understand this impact cited that although concerns of contracting the disease were prevalent, just as prevalent if not more in some cases was fear and anxiety over economic factors, general uncertainty, loss of social support groups, and psychological impacts of social isolation among others (Brooks et al., 2020; Cowan et al., 2020).

A key aspect of the pandemic was that its impact was not homogenous. Considering differences between individual circumstances and different government responses and timelines it stands to reason that the severity and manner of impact would also be different. Studies on the impact of the pandemic that focus on socio-economic and demographic factors align with this thought process. They find differences in the stressors between men and women, the young and elderly, between minority groups and others, and many other subgroups (Pedrosa et al., 2020; Qiu et al., 2020; Zhang et al., 2020). These differences lead to an important question; how are people affected by COVID-19? Answering this question is a crucial first step in revealing key areas where we need to improve our ability to safeguard individuals' physical and mental wellbeing. Even if we are not faced with another pandemic of this scale, there could be otherwise unnoticed or underrepresented issues that minority groups are faced with.

Literature Review

Analysis of social media data is becoming increasingly popular as data methods evolve and the amount of publicly available data grows. Twitter, perhaps due to its popularity and also format of short messages is one platform that has seen use in numerous different academic studies (Ortiz-Ospina, 2019). The applications for twitter data include linguistics, criminology, disaster management, event prediction, human mobility, and many others (Pradyumn, 2018). In the field of public health research specifically, Twitter has been used in event detection, pharmacovigilance, surveillance, forecasting, disease tracking, and geographic identification to name a few broad categories (Edo-Osagie et al., 2020).

From even a brief literature review it is evident that Twitter has great usefulness in both academic research and real-world applications. The pairing of Twitter data with natural language processing methods is an especially popular application, and the specific use of natural language processing on Twitter data to explore the impact of COVID-19 is itself not a new idea. However, the techniques that are employed from study to study vary based on the focus of the research. Techniques used in previous research commonly include sentiment analysis and topic modeling such as Latent Dirichlet Allocation (LDA). Many studies that employ these methods are also limited in scope to a single country (Marshall et al., 2022; Sengupta et al., 2020).

Data and Methods

As a result of changes to the Twitter API the ideal option of pulling data relevant to this study was not available. Searching for publicly available datasets of tweets relevant to this study also yielded no results. Therefore, the data employed in this study was taken from one of my previous projects where twitter data had been retrieved before changes to the API. The time periods for the retrieved tweets aligned close enough to the 3 periods required for this study to be applicable. The periods ranged from March 2018 to October 2018, September 2019 to July 2020, and September 2021 to July 2022. These represent a period before the pandemic, the start of the pandemic, and a period after the start of the pandemic. The dataset consists of about 480,000 tweets across all 3 periods. Because period 1 is slightly shorter it contains about 100,000 tweets while periods 2 and 3 contain about 180,000 tweets each. The tweets contain a sample of users from the Dallas Forth Worth (DFW) Metroplex as this was the area of interest for the original study the tweets were used in. A single bot account was removed as it far outranked all other users in number of posts, and on further inspection the account stated that it was used for automated posting. Due to hardware limitations, I took a random sample of 10,000 tweets from each period, about the maximum number of tweets that could be processed at a time with hardware I had access to. No other filtering was performed on the data.

The natural language processing methods employed in this study are topic modeling and a dictionary method. I preprocessed the tweets by removing punctuation, numbers, stop words, and whitespace, transforming all words to lowercase, and stemming terms. Applying topic modeling on the tweets allows us to examine how the pandemic affected individuals by grouping together the words most commonly associated with each other across tweets. This is a powerful unsupervised machine learning technique that can serve as an alternative to traditional survey methods. It has the advantage of not leading responses with questions like a survey might, but has a disadvantage of information loss with less frequently mentioned topics (Grimmer et al., 2022; Roberts et al., 2014). A dictionary method on the other hand is useful in observing the intensity of mental health issues that occurred in my sample. It consists of a simple count of words across all tweets that are found in a supplied dictionary. The ideal dictionary for this study would include terms that only capture discussions related to mental health. However, the simplicity of the dictionary method means that complex ideas such as mental health, which can include numerous terms used commonly in other topics, may not be well captured (Grimmer et al., 2022).

Table 1. Mental Health Dictionary

depress, anxiety, anxious,
stress, strain, frustrat,
fear, frighten,
alone, lone,
scari, scare, scared, scary,
danger, racist, racism, mental

The dictionary used in this study, seen in Table 1, was built based on a literature review of previous mental health studies using Twitter data and an analysis of the terms used in this study's collected tweets.

Results

The results from topic modeling the tweets can be seen in Figures 1, 2, and 3, representing periods 1, 2, and 3 respectively. These topic models are clear indicators of one of the primary problems in this study, the overpowering noise present in an unfiltered twitter query. Even amidst a major global event like the pandemic there are no topics that can be inferred to relate to COVID-19, the pandemic, or mental health in a reasonable way. All the modeled topics are plagued by the amount of noise, with a clear topic label not able to be inferred easily. I attempted to reduce noise through a method in the 'stm' package, setting a minimum threshold for how many times a term must appear across all documents to be kept in the model. Terms that do not meet the minimum threshold are dropped. A higher threshold reduces noise in the topics but at the cost of also reducing information. However, adjusting this threshold did not produce more coherent topics.

Figure 1. Period 1 Topic Model

Topic 1:
know, love, thank, lol, got, day, need, can, que, want, lmao, say, realli, peopl, game, great, year, never, even, man

Topic 2:
unicaesradio, accident, del, dalla, carril, tráfico, los, nuevo, alerta, amp, call, great, way, drink, day, got, need, alway, back, good

Topic 3:
ephoustonbil, johanbbt, zappafay, cheer, jonmontag, dregofish, trafico, casaskullmark, manvsal, rjellyman, justbeerlov, jagoff, beerguypdx, thestraighthop, jacobgrim, wolvb, paulthebeerguy, larryburnett, mtravi, jwagsjack

Topic 4:
one, now, back, work, right, well, today, ’re, way, best, drink, alway, got, year, ’ve, call, thing, good, make, see

Topic 5:
just, get, like, good, will, time, make, amp, don’t, look, think, see, come, shit, still, drink, take, feel, thing, new

Figure 2. Period 2 Topic Model

Topic 1:
make, accident, better, peopl, 've, love, need, thank, got, cheer, one, know, new, will, happi, day, say, way, come, think

Topic 2:
cellmavin, jonmontag, johanbbt, ephoustonbil, cedarflat, lhtetrick, rjellyman, tammyjaben, zappafay, justbeerlov, wolvb, dregofish, okitim, qblacklock, manvsal, thestraighthop, jacobgrim, jwagsjack, bplohocki, realbmaxwel

Topic 3:
lol, nigga, check, still, shit, item, 'll, tri, like, say, one, start, play, can't, realli, just, good, watch, look, love

Topic 4:
get, good, time, will, look, think, amp, great, man, well, shop, can't, cheer, like, one, tri, say, back, come, thing

Topic 5:
like, just, one, don't, know, thank, got, want, day, need, can, love, see, now, realli, work, peopl, say, back, come

Figure 3. Period 3 Topic Model

Topic 1:
que, watch, can't, see, can, get, one, come, work, game, best, much, never, make, amp, take, let, just, yes, time

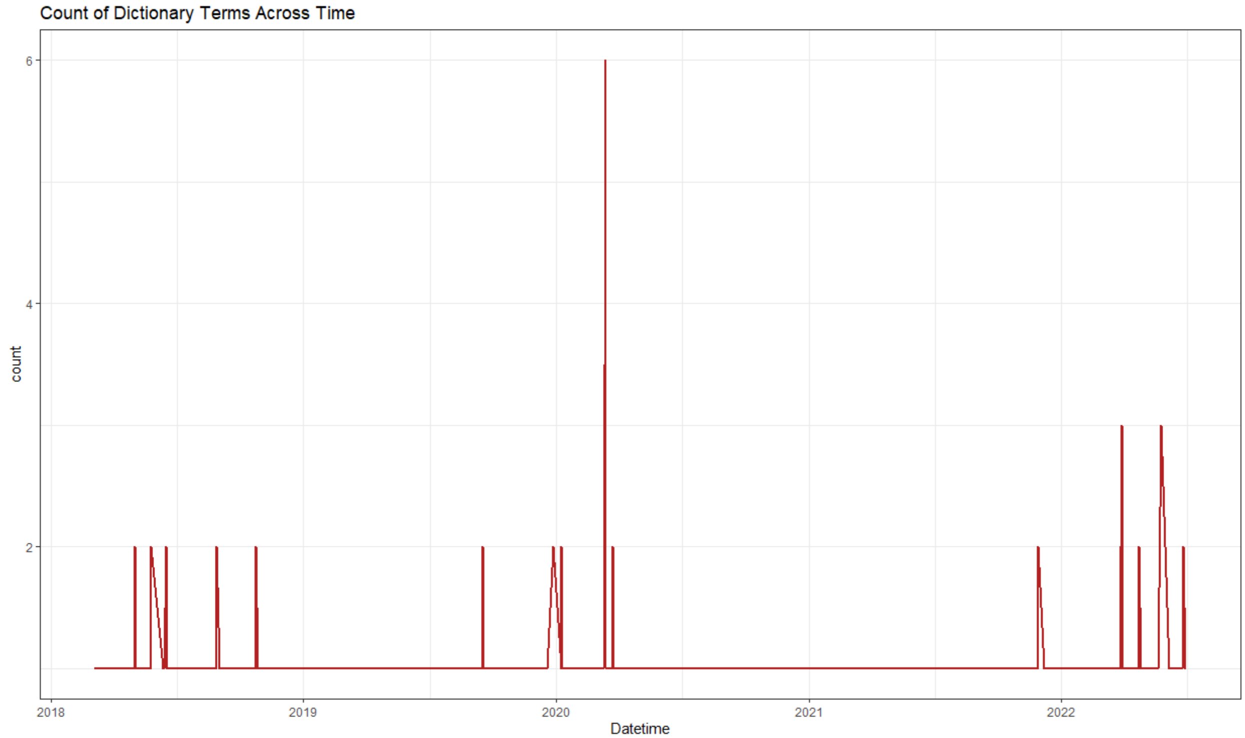
Topic 2:
johanbbt, jonmontag, cellmavin, ephoustonbil, justbeerlov, rjellyman, tammyjaben, manvsal, realbmaxwel, beerhunt, senorgreezi, cheer, qblacklock, dregofish, kubrickx, amethystheel, catbrew, great, good, damn

Topic 3:
thank, love, good, time, day, one, will, amp, back, come, work, great, year, happi, game, take, play, still, well, let

Topic 4:
just, like, get, lol, need, can, don't, look, make, see, realli, right, man, never, feel, yes, lmao, alway, damn, let

Topic 5:
know, now, got, want, think, peopl, say, today, way, thing, texa, see, can, much, one, just, right, get, work, don't

Figure 4. Count of Dictionary Terms Across Time



The results from the dictionary method can be seen in Figure 4, a graph of the count of dictionary terms across all tweets over time. The amount of noise present in the corpus of tweets is evident here also. A sample of 30,000 tweets across all 3 periods yielded a maximum count of 6 terms and a mean count of about 1 term. Although we see a spike in term count at the start of the pandemic as we would expect, it subsides quickly. With this small sample size we can't confidently confirm trends regarding the effect of the pandemic on the intensity of mental health issues.

Model Validation

A critical step in any modeling process is validating the results. In this study, validating the dictionary method means ensuring that the model is truly capturing discussions on mental health. There is some error because the model captures tweets not related to mental health while failing to capture those that actually are related but don't contain any of the dictionary terms. This error is normal but should be minimized. For this purpose, I manually analyzed the tweets that the model captured to assess terms that are capturing too many unrelated tweets. I removed and added terms while testing model validity until I found an acceptable balance between minimizing error and maximizing the amount of mental health tweets captured. There is also an assumption being made here that an increase in mental health terms corresponds to worsening mental health among individuals. This assumption is strengthened by including a baseline in the form of the period before the pandemic and a period after the height of the pandemic.

Validating the topic model consists of defining coherent topic labels. It is important to consider the context of the terms in the topic when doing so and construct a clear label that encompasses all terms well. Topics that are not coherent lack validity. In this study, the noise in the corpus results in topics that are largely incoherent and therefore lack validity. In other words, they cannot be applied with confidence to answer our question of how COVID-19 affects people.

Conclusion

How are people affected by COVID-19? Differences between individuals and states in their capabilities and responses to COVID-19 make this a question that is best answered through a survey. As an alternative method to traditional surveys, I employ natural language processing techniques including topic modeling and a dictionary method to tweets. Although there was too much noise present in my corpus of tweets to confidently answer my question, I highlight an important aspect of applying the methods to twitter data. To avoid detrimental noise, the tweets should be filtered when they are pulled from Twitter, which is possible through the Twitter API. If you supply a list of topics the API will only pull tweets that are labeled with those topics. This will allow you to curate a significantly less noisy corpus.

References

Brooks, S. K., Webster, R. K., Smith, L. E., Woodland, L., Wessely, S., Greenberg, N., & Rubin, G. J. (2020). The psychological impact of quarantine and how to reduce it: rapid review of the evidence. The lancet, 395(10227), 912-920.

Cowan, K. (2020). Survey results: Understanding people's concerns about the mental health impacts of the COVID-19 pandemic. MQ: Transforming mental health and the academy of medical Sciences, 2020.

Edo-Osagie, O., De La Iglesia, B., Lake, I., & Edeghere, O. (2020). A scoping review of the use of Twitter for public health research. *Computers in biology and medicine*, *122*, 103770. https://doi.org/10.1016/j.compbiomed.2020.103770

Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. (2022). Text as Data: A New Framework for Machine Learning and the Social Sciences. Princeton University Press. Chapter 13.

Marshall, C., Lanyi, K., Green, R., Wilkins, G. C., Pearson, F., & Craig, D. (2022). Using Natural Language Processing to Explore Mental Health Insights From UK Tweets During the COVID-19 Pandemic: Infodemiology Study. *JMIR infodemiology*, *2*(1), e32449. https://doi.org/10.2196/32449

Ortiz-Ospina, E. (2019). The rise of social media. OurWorldInData.org. https://ourworldindata.org/rise-of-social-media

Pedrosa, A. L., Bitencourt, L., Fróes, A. C. F., Cazumbá, M. L. B., Campos, R. G. B., de Brito, S. B. C. S., & Simões E Silva, A. C. (2020). Emotional, Behavioral, and Psychological Impact of the COVID-19 Pandemic. *Frontiers in psychology*, *11*, 566212. https://doi.org/10.3389/fpsyg.2020.566212

Pradyumn, M., Kapoor, A., & Tabrizi, N. (2018). Big Data Analytics on Twitter: A Systematic Review of Applications and Methods. *Big Data – BigData 2018*, 326–333. https://doi.org/10.1007/978-3-319-94301-5_26

Qiu, J., Shen, B., Zhao, M., Wang, Z., Xie, B., & Xu, Y. (2020). A nationwide survey of psychological distress among Chinese people in the COVID-19 epidemic: implications and policy recommendations. *General psychiatry*, *33*(2), e100213. https://doi.org/10.1136/gpsych-2020-100213

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, David G. Rand. (2014). "Structural Topic Models for Open-Ended Survey Responses." American Journal of Political Science 58: 1064-1082.

Sengupta, S., Mugde, S., & Sharma, G. (2020). An Exploration of Impact of COVID 19 on mental health -Analysis of tweets using Natural Language Processing techniques. medRxiv. https://doi.org/10.1101/2020.07.30.20165571

Siddiqui, S., Alhamdi, H. W. S., & Alghamdi, H. A. (2022). Recent Chronology of COVID-19 Pandemic. Frontiers in public health, 10, 778037. https://doi.org/10.3389/fpubh.2022.778037

Zhang, J., Lu, H., Zeng, H., Zhang, S., Du, Q., Jiang, T., & Du, B. (2020). The differential psychological distress of populations affected by the COVID-19 pandemic. *Brain, behavior, and immunity*, *87*, 49–50. https://doi.org/10.1016/j.bbi.2020.04.031