12-2022

Mining Public Opinion on COVID-19 Vaccines using Unstructured Social Media Data

Chad Aaron Melton
cmelton3@vols.utk.edu

To the Graduate Council:

I am submitting herewith a dissertation written by Chad Aaron Melton entitled "Mining Public Opinion on COVID-19 Vaccines using Unstructured Social Media Data." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Data Science and Engineering.

Arash Shaban-Nejad, Major Professor

We have read this dissertation and recommend its acceptance:

Drs Kathleen Brown, Chuanren Liu, Eun Kyong Shin

Accepted for the Council:

Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Mining Public Opinion on COVID-19 Vaccines using Unstructured Social Media Data

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Chad A. Melton

December 2022

DEDICATION

For Alison and Milo…

ACKNOWLEDGEMENTS

First and foremost, I would like to thank Arash Shaban-Nejad for opportunity to work under his guidance and support. I could not have hoped for a better advisor and mentor. I would like to thank my committee – Drs. Kathleen Brown, Chuanren Liu, and Eun Kyong Shin for their superb guidance and for helping make this research much better. This work would not be possible without your input. I would like to sincerely thank the National Cancer Institute for funding my research. I would also like to thank past and current members of the CMBI research group at UT Health Science Center, and the Bredesen Center staff from the University of Tennessee for their knowledge, support, and the opportunity to pursue this path. I would like to offer sincere appreciation to Brianna White, Olufunto Olusanya, Paria Zareie, ZareWhitney Brakefield, Lokesh Chinthala, Nariman Ammar, Akram Mohammed, Rebecca Brink, Kevin De Angeli, and Peyton Vanhook. Nicole Siuti , Charinjit Chaturlal, Budhu Bhaduri, David Page, Joshua Emery, April Morton, Melissa Dumas, Mark Littman, Russell Zaretzki, Audrey Martin, Nathanial Raymond and the Y'all Real Crew, Henry Simpson, Dibyendu Mukherjee, BC 2019 cohort, Subhadeep Chakraborty, Salman Khan, James Sousa, Mike Seal, Carl Sagan, and Richard Berendzen. This work would not have been possible without your inspiration, friendship, lessons, and guidance. I would like to thank my family and grandparents for their encouragement and providing interesting experiences that made me curious about the universe that we find ourselves a part of. Finally, I would like to thank my wonderful wife Alison (and more recently my son Milo) for hanging in there with me over the last 8 years with high spirits. I am forever grateful for your love and support.

ABSTRACT

The emergence of the novel coronavirus (COVID-19), and the necessary separation of populations led to an unprecedented number of new social media users seeking information related to the pandemic. Nowadays, with an estimated 4.5 billion users worldwide, social media data offer an opportunity for near real-time analysis of large bodies of text related to disease outbreaks and vaccination. This study investigated and compared public discourse related to COVID-19 vaccines expressed on two popular social media platforms, Reddit and Twitter. Approximately 9.5 million Tweets and 70 thousand Reddit comments were analyzed from dates January 1, 2020, to March 1, 2022, and analyzed through topic modeling, sentiment analysis, and semantic network analysis. Sentiment analysis through the fine-tuned DistilRoBERTa model revealed that even though Twitter content was overall more negative than content expressed on Reddit, relatively similar changes in sentiment occurred among users of both online platforms. Reversals in sentiment trends typically occurred within relative proximity to events such as vaccine development news, vaccine release, frequent discussion of side-effects, the discovery of new variants, and pandemic fatigue. Topic modeling and semantic network analysis provided insight into how public discourse related to COVID-19 and vaccinations, misinformation, and vaccine hesitancy evolved over 26 months. Though misinformation and mention of conspiracy theories were detected with the analysis, the occurrence of both was less frequent than expected. This work provides a framework that could be scaled and utilized by public health officials to monitor disease outbreaks in near real-time in large communities as well as smaller local groups. Hopefully, the results from this study will help to guide and facilitate the implementation of targeted digital interventions among vaccine-hesitant populations and provide insights to public health officials to inform decision-making and effective policy development.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I: INTRODUCTION

Motivation and Background

In late December of 2019, the highly transmittable coronavirus disease 2019 (COVID-19) acquired through the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-COV-2), began its rampage impacting every aspect of life throughout all societies of the world. COVID-19 was declared a pandemic by the World Health Organization (WHO) in March 2020, and nearly a year later, approximately 150 million individuals had been infected (confirmed) and 2.8 million have died (WHO, 2021). Vaccines are determined to be one of the most effective interventions for preventing and controlling the spread of the COVID-19 pandemic. Although the accelerated development of vaccines was unprecedented, improving public sentiments for vaccine uptake and diffusing widespread skepticism towards science has been extremely challenging, particularly against the backdrop of the COVID-19 pandemic (WHO, 2014). This fact is clearly illustrated by the unwillingness and reluctance among certain populations across the world to be vaccinated against Covid-19 (WHO, 2014).

Vaccine hesitancy, classified among the top ten threats to global health by the WHO, is defined as the “delay in acceptance or refusal of vaccines despite availability of vaccine services” (WHO, 2014). Vaccine hesitancy/refusal/delay is perceived to originate from a diverse, multifaceted, and often concurring array of underlying factors ranging from religion, political ideology, and the anti-vaccination movement, to outlandish conspiracy theories and beliefs (Rutjens et al., 2021). Current drivers for COVID-19 vaccine hesitancy include disinformation, misinformation, conspiracy beliefs propagated through social media, inadequate and contradictory responses from the federal government, frustrations among the general public, and fear of the unknown (Puri et al., 2020). Apprehension regarding the vaccine’s safety, side effects, efficacy, and access also contribute to vaccine hesitancy. For instance, the recent pause in the roll-out of the Johnson & Johnson's Janssen (J&J/Janssen) COVID-19 vaccine due to reported rare side effects from blood clots has reignited fears regarding vaccine uptake.

Exacerbating the COVID-19 related-health messaging crises are the sensational design of vaccine-related misinformation and "fake news" which have tended to spread more rapidly than factual evidence-based information (Balmas 2014). Concernedly, this spread of vaccine disinformation and misinformation ultimately leads to quantifiable negative outcomes (e.g., low vaccination rates, increasing hospitalization rates, morbidity, and mortality from vaccine-preventable diseases, etc.) (Zhang et al., 2021, van der Linden et al., 2021). Needless to say, the COVID-19 vaccine roll-out has been challenging due to vaccine hesitancy/delay/refusal thus the urgent need for a call to action.

Objectives

As of August 28, 2022, the United States has administered about 609 million doses of vaccine. Additionally, a total of 12.6 billion doses of vaccine have been administered worldwide (including in the U.S) (WHO, 2022). Despite the nebulous onslaught of disinformation and misinformation regarding the COVID-19 vaccine, results of recent surveys suggest that public opinion/confidence is improving and has increased from approximately 51–69% from September 2020 to March 2021 (Funk and Tyson 2020, Funk and Tyson 2021). This slight increase in positive views regarding the COVID-19 vaccine is one step in the right direction to improving vaccination acceptance. However, current vaccination rates are still significantly below the percentage threshold required in the U.S. (i.e., 79–90%) (Aschwanden 2020). Therefore, it is increasingly imperative to assess public sentiment and to understand what drives vaccine hesitancy.

Sentiment analysis and topic modeling are analytical tools that can be utilized to systematically identify, extract and measure the sentiment and topics from subjective textual data (e.g., obtained from social media posts, customer reviews, online survey responses, etc.) quickly, effectively, and inexpensively, as well as extrapolate common themes throughout a document. In a step in that direction, the major objectives of my research are to 1) detect misinformation/disinformation and 2) document how misinformation has diffused throughout the pandemic. These objectives fall into two analytical categories, a) sentiment analysis and b) semantic analysis through topic modeling and graph network analysis.

Due to recent moderately positive polling results (Funk and Tyson 2020), one goal of this work is to investigate the sentiments reflected in discussions related to the COVID-19 vaccines throughout our data set and determine whether these sentiments reflect the overall public sentiment towards vaccinations reported in polling. With this portion of the analysis, I expected to see changes in sentiment over time, as well as sentiment to be more negative than positive on average. Additionally, this work explored public discourse regarding COVID-19 vaccination expressed on two social media platforms, Reddit and Twitter. Through topic modeling and semantic network analysis, I expected to detect evidence of vaccine hesitancy, and discussions of conspiracy theories. Furthermore, I expected to see evidence of social media users' discussion of COVID-19 symptoms, public health measures, and experiences with vaccination. Stemming from the main objectives of my research, an additional goal is to compare the results of these analytical methods from two social media platforms, the Extended Reddit data (see 4.1.2) and *Coronavirus Twitter Data*. This supplementary analysis provides insights into the similarities and differences between two thriving social media users. Based on functionality and demographic information, I expect major differences in discussion topics between Reddit and Twitter. Lastly, I expect the sentiment expressed on these two platforms to be similarly negative. The methodology employed to achieve these objectives facilitates the comparison of two vibrant and popular social media platforms. Lastly, this work provides a framework that could be scaled and utilized by public health officials to monitor disease outbreaks in near real-time.

Approach

With public health in mind, the objectives of this research were achieved by utilizing a combination of computational and analytical tools to investigate sentiment and topic modeling. This study harvested approximately 100k Reddit posts and approximately 13 million Tweets related to COVID-19 vaccination and vaccines. Sentiment analysis was conducted in two different ways including the use of a Python package (*Textblob*) and a state-of-the-art language model that was fine-tuned by training a DistilRoBERTa model with my own labeled data set. The investigation of public discourse on Reddit and Twitter text data was conducted with the popular topic modeling method, Latent Dirichlet Allocation. This portion of my work also employed the building of semantic networks to record betweenness centrality, transitivity, and other network statistics. The networks also served as a qualitative aid in interpreting and visualizing discussion topics.

Research Questions

To test my hypothesis, I posed the following research questions:

1. How has sentiment changed throughout the pandemic?
2. Is it possible to correlate changes in sentiment with major events during the pandemic?
3. How does sentiment vary between Reddit communities and Twitter?
4. Do Reddit and Twitter discuss similar topics at similar times?
5. What are commonly discussed topics on Reddit and Twitter related to COVID-19 vaccines?
6. Will the occurrence of misinformation be prevalent in data?
7. Will the occurrence of conspiracy theories be prevalent in data?
8. Is one social media platform better for public health surveillance?
9. How generalizable is the approach presented in this research?

Contribution

During these dynamic and challenging times, I anticipate that this study will offer insight into the general public's sentiments/opinions regarding the COVID-19 vaccine using a relatively unexplored dataset. The goal of the research to be presented is to surveil public discourse related to COVID-19 vaccination from large corpora of social media textual data through sentiment analysis, topic modeling, and semantic network analysis. The sentiment analysis of discussion on two major popular social media platforms sheds light on possible differences in users' experiences during the COVID-19 pandemic. Temporal LDA topic modeling and semantic network analysis provides insight into how public discourse related to COVID-19 and vaccinations, misinformation, and vaccine hesitancy evolved over 26 months. This work also offers two large datasets (*Coronavirus Twitter Data and Reddit 2.0 Reddit data*) consisting of over 13.5 million text entries that can be utilized by researchers to further investigate phenomena related to vaccine perception and public discourse during one of the worst pandemics in recent history. Furthermore, this work provides a framework that could be scaled and utilized by public health officials to monitor disease outbreaks in near real-time. Hopefully, the results from this study will help to guide and facilitate the implementation of targeted digital interventions among vaccine-hesitant populations and provide insights to public health officials to inform decision-making and effective policy development.

CHAPTER II: LITERATURE REVIEW AND RELATED WORKS

Natural Language Processing and Disease Surveillance

Introduction

Since the invention of the internet and the proliferation of social media platforms, *Natural Language Processing* (NLP) techniques have been used successfully across many domains. Early social media platforms such as Prodigy and Compuserve attracted the attention of researchers to understand and surveil discussions of various social phenomena. As societies explored this new technology, and internet culture began to materialize, researchers quickly utilized this new abundance of textual data (Sproull and Faraj, 1997). Public health researchers were some of the earliest to see the value of this new technology where users established focused online communities to gather information, establish friendships, and express their feelings, and sentiments. Nowadays, advancements in NLP have made it possible for near-real time monitoring systems to be developed to understand public health events, monitor discussions regarding public health, and even predict disease outbreaks. This section offers an overview of several recent studies that are explicitly relevant regarding public health surveillance in concert with the computational techniques utilized for this body of work.

NLP Methodologies

Sentiment Analysis

Sentiment analysis is the practice of extrapolating the sentiment of a subject, idea, event, or phenomena by classification of written texts as some value of polarity (i.e., positive, negative, or neutral) (Liu, 2012). Because gauging public sentiment is vastly important to determining appropriate messaging, intervention, and policies, this sentiment analysis, and topic modeling techniques have been used in many scientific, social, and commercial applications. Sentiment analysis of social media is a relatively new field. Some early works analyzed Twitter data to detect sentiment in product reviews to inform potential consumers while others questioned and tested whether microblogs (such as Twitter) were better for sentiment analysis than longer documents (Go et al., 2009., Bermingham and Smeaton, 2010). Other studies employed sentiment analysis techniques to gain insight into political elections (Wang et al., 2012) to monitor public opinion around the world regarding mask-wearing during the pandemic (Sanders et al., 2020).

Topic Modeling

Topic modeling has a rich history of versatile use cases. Techniques such as Latent Dirichlet Allocation (LDA) excel at providing a statistical and probabilistic overview of topics that are present in large corpora of texts. Similar to sentiment analysis, the technique is regularly employed by industry to investigate the public perception of various products (i.e., restaurants, cars, online shopping, etc.). LDA assumes that documents with similar topics will use similar diction and that the topics will display a sparse Dirichlet distribution. For example, every word in the document is randomly assigned to a user-defined number of topics T. The algorithm then calculates the proportion of words in each document assigned to a topic (i.e., [p(topic T | document D)])and then the proportion of words that were assigned to a topic overall document (i.e., [p(word W | topic T)]). The product of these pro-portions is computed for each topic T and compared to every other topic T until algorithmic convergence is achieved (Srinivasa-Desikan, 2018., Gensim, 2011., Blei et al., 2003). In the public health domain, researchers have employed topic modeling to observe trends in public health messaging (Park and Park 2021). Topic modeling has also been used to research health disparities from Twitter posts or tweets (Mantas, 2020), and obesity (Ghosh and Guha, 2020). More recently, studies have attempted to use topic modeling to predict vaccine hesitancy (Krishnan et al., 2021).

Network Analysis

Networks analysis has been employed to understand problems for many centuries. One of the earliest documented studies dates back to the 1730s when Euler used topology and graph network theory to provide solutions to the "Seven Bridges of Konigsberg" problem (Euler 1736). Network analysis has a rich history of recognizing structures in large networks (Clauset et al., 2004) and has equally been used across many scientific domains, especially in social media and influence networks. More recently, computational semantic analysis has been used to explore clustering in news articles and Google reviews (Verimyev et al., 2019) as well as educational textbooks (Srinivasa-Desikan, 2018). More current work has been involved in the public health domain where a study by Shin et al investigated comorbidity networks of Korean hospital patients infected with COVID-19 (Shin et al., 2021). Others focused more specifically o COVID-19-related networks based on specific geographic locations by languages (i.e., Italian Twitter, China) (Mattei et al., 2021, Gao et al., 2021). Finally, a study used semantic network analysis to investigate COVID-19 emotions displayed in news media (Yoo et al., 2021).

Bidirectional Encoder Representations from Transformers

Significant advances in natural language processing have occurred since the development and work built from the architecture of Bidirectional Encoder

Representations from Transformers (BERT). The development of BERT and variations of BERT have essentially propelled NLP into a new era (e.g., stone age to bronze age). BERT is a powerful and versatile AI-based natural language processing algorithm developed at Google AI Language that excels at text classification (*i.e., ontologies, categories, sentiment, etc.*) of unstructured/semi-structured text data characteristic of social media data (Devlin et al., 2018). BERT algorithm was trained on the entirety of *Wikipedia* and the *Brown Corpus* over four days of 16 cloud-based TPUs (Tensor Processing Units). BERT is a transformer-based language model that employs multiple encoders to create word embeddings. These embeddings are then used in concert with masked language modeling (MLM) and next sentence prediction (NSP) to learn by predicting random masked words in a sentence and then learning to predict sentences respectively. These two steps teach BERT to understand context, a skill that older recurrent neural networks typically struggled with. A convenient aspect of BERT is that it has the capability to fine-tune the model with relevant data by replacing the output layer with weights from custom data. Researchers have been inspired by the original BERT (OB) architecture to create many variations (e.g, RoBERTa, DistilRoBERTa, DistilBERT, BART, etc) that have surpassed the benchmarks of previous models. Moreover, these models can be fine-tuned for specific domain-based tasks.

Applications of these models are extensive and their implications are far-reaching. NLP community repositories such as Huggingface.co have provided members of the NLP community with hundreds of pretrained models. As long as the NLP researcher has a bit of training data, these models can be adapted these employed to extract sentiment (Melton et al., 2021), emotion (Jeon et al., 2022), clinical note categorization (ClinicalBert, BioBERT) (Alsentzer et al., 2019, Lee et al., 2020), and topic classification across various languages (Xenouleas et al., 2022). BERT and variations of BERT are the current state-of-the-art in natural language processing.

Public Health Surveillance and Disease Outbreak

Alessa and Faezipour (2019) applied multimodal analytical techniques to track disease outbreaks and develop early warning systems (Alessa and Faezipour., 2019). Their study used linear regression, sentiment and content classification, and mapping for theirs. The study used an amalgamation of previously collected data (Lamb et al., 2013, Sanders 2015), approximately 8.4 million Tweets harvested themselves, and from the self-reporting influenza tracking site, Flu Near You (Chunara et al., 2013). The Lamb et al and Sanders data were used to train a FastText classifier to determine if a Tweet was related to influenza or unrelated and employed the Python library, TextBlob to classify sentiment. To demonstrate the effectiveness of their pipeline, the researchers also test additional classifiers including SVM, KNN, Naïve Bays, decision trees, random forests, and AdaBoost. Mapping was completed with MapReduce, a module that assists in organizing data for parallel processing. The results of their FastText classifier reported “accurate” results with an F-measure of 89.9 percent while their regression estimator reported a 96.29% correlation with data obtained from the CDC. Alessa and Faezipour offered a novel and multimodal approach to predicting weekly influenza infection rates. This study excelled in evaluating the performance of their model against other proven machine learning and regression methodology. The authors also provided an excellent description of background work and statistical analysis and evaluation. However, this article could have been a bit more thorough with consistent reporting of the results (e.g., lack of sentiment analysis results).

Inspired by the rise of medical misinformation, a study by *Raghupathi et al* used a combination of clustering and sentiment analysis to investigate approximately 9500 Tweets posted during an outbreak of the measles virus in the United States during the first half of 2019. Topic clustering in this study was achieved in several steps, first through a vectorization algorithm, *term-frequency-inverse document frequency* (TF-IDF). TF-IDF is conducted by calculating the product of two metrics: the frequency of a term and the inverse document frequency.

As an additional step, the K-Means clustering algorithm was then used to further classify Tweets. The K-Means algorithm is a simple unsupervised clustering algorithm that works by iteratively randomly placing a user-determined number of clusters of centroids. Data points are then assigned to each centroid. Centroid locations are then recalculated based on distance (usually Euclidian) to data points. Once centroid locations stabilize, the algorithm is finished. It should be noted that the K-Means user should have some previous knowledge about their data set so the analysis is properly constrained. Problems in analysis can occur if the number of clusters is not estimated correctly. K-Means has been used successfully in a wide variety of scientific domains including social network analysis, spectroscopy, GIS applications, gaze detection, and many others (Likas et al., 2003). For the sentiment, this study utilized the Valence Aware Dictionary and Sentiment Reasoner (VADER) sentiment classifier, an algorithm within the larger Natural Language ToolKit

(NLTK) Python package. Classification results of this work revealed that approximately 77 percent of Tweets focused on the need for vaccines to combat diseases such as Ebola, HPV, and influenza. Approximately 50 percent focused on the measles outbreak, while the remaining 50 percent were focused on debates between vaccine supporters and anti-vaccine users. Of these data, 40.3 percent were classified as positive, 43.4 percent negative, and 16.3 percent neutral. Interestingly, this study found a minimal discussion of misinformation within their data. Though the methods in this study are not state-of-the-art, this study is significant in that the authors were motivated by the negative effects that misinformation has in the public health domain, especially related to vaccine hesitancy.

While this subsection describes two examples in detail of work that offered significant contributions to broad public health surveillance regarding disease outbreaks, many other researchers have provided excellent work regarding social media and disease outbreaks. In a systematic review by Tang et al., 2018, the authors document at least 15 studies related to H1N1 (swine flu), 10 related to Ebola, H7N9 (avian flu), two related to West Nile, and one for each EHEC, MERS-CoV, and measles. Similar techniques have continued to be applied during the times of the COVID-19 pandemic. A study by Gbashi et al. (2021) focused on detecting the opinion of media polarity on the COVID-19 vaccine in Africa with Twitter and Google News articles. Furthermore, additional research investigated public sentiment in India (Praveen et al., 2021, Dubey 2021), Indonesia (Ritonga et al., 2021), China (Yin et al., 2021), and North America ((Jang et al., 2021).

NLP Approaches for Studying the COVID-19 Pandemic

Wu et al published a study regarding public sentiment towards the COVID-19 vaccine and topic modeling of several subreddits (Wu et al., 2021). The authors of this study harvested 57657 posts from eight subreddit communities, two of which were directly related to COVID-19 (r/COVID-19 and r/Coronavirus. The remaining subreddits included were from a wide variety of topics not specifically focused on the virus (r/worldnews, r/conspiracy, r/politics, r/wallstreetbets, r/AskReddit, and r/news). After cleaning the text data with NLTK, the authors employed LDA to model latent topics within the corpus. The authors also use Linguistic Inquiry and Word Count (LIWC) with hopes to capture "psychological" information from the text. Their LDA analysis concluded that 10 topics were optimal. The authors used words in each topic to manually assign broader topics (e.g., Skeptical/aggressive remarks, Life/family/kids, Stockmarket/sports, etc.). The authors then provide a linguistic profile for text composed by users across three different subreddits. We et al have many merits and the design of the project was superb. The study excelled with the usage of LDA and provided excellent visual aids to display which subreddits related to each topic. This work also provided superb visualization for the linguistic profiles of users who are active in multiple subreddits. However, the author's interpretation leads them to declare that "Reddit has served as a hotbed for conspiracy and misinformation". And it is agreeable that some of the subreddits are filled with misinformation but the authors also fail to realize the potential for echo chambers that can occur in some Reddit communities and fail to mention the potential for bias in their chosen data. It is reasonable to include r/conspiracy in the context of COVID-19 because misinformation has been a significant problem around public health issues. Nonetheless, these authors provide no real explanation as to the reason behind their choice of subreddits that have no real relevance to the COVID-19 pandemic (e.g., r/wallstreetbets). Furthermore, the authors did not appear to have queried their data for terms related to the pandemic, hence the detection of topics related to stocks and sports.

Luo et al compared perceptions of COVID-19 vaccines on social media between the United States and China with Twitter and Weibo data respectively (Luo et al., 2020). The authors used semantic network analysis to explore the interconnectedness of terms and topics in these two online communities that were posted between December 1, 2020, and February 20, 2021. Approximately 750000 Tweets and 360000 thousand Weibo posts were harvested and filtered for specific terms related to COVID-19 vaccination, and only the original messages were included (i.e., no retweets or shares).

The authors followed traditional text cleaning methods (i.e. tokenization and lemmatization) before constructing word co-occurrence matrices for each corpus. These matrices were then used to construct semantic networks with the network building software, Gelphi. Network density, degree, and eigenvector centrality (EC) were then calculated. Density represents the ratio of total present edges to the number of total possible edges. The degree in a network represents the number of edges that are connected to each word. EC measures the influence of a node on the entire network. Sentiment analysis was also conducted on Tweets with the LIWC while the sentiment of Chinese texts was calculated with TextMind, a tool from the Chinese Academy of Sciences. The EC and degree results show that some similarities exist between the top 10 words for each corpus, mainly in terms relating to COVID-19 and vaccines, or vaccination. Overall, the Twitter data were more focused on vaccination and health while the Weibo data included more terms related to geography (e.g., China, America, country, and global). Typical EC values in the Weibo data set were generally higher than values in the Twitter data. For the Twitter data, 49.99 percent of the Tweets were neutral, 30.62 percent were positive, and 19.40 percent were negative. 40.64 percent of the Chinese posts were positive, 37.44 were neutral, and 21.92 were negative. These results also show one of the major differences between the two datasets exists in discussions about taking the vaccine. Discussions about personal experience with taking the vaccine were much more prevalent with Twitter data than with Weibo.

This study provides an outstanding usage of network analysis and sentiment analysis to investigate unique societies that could be considered opposites in nature (i.e., collectivist and individualist). Moreover, the author states that due to the collectivism in Chinese society, many Chinese citizens often do not reveal personal feelings regarding such matters. It is highly conceivable that the true feelings of the Weibo community could also be withheld or filtered by Chinese governmental surveillance systems. Thus, a true comparison between the two might be impossible to accurately conduct.

Kim et al embraced the power of several different sized BERT models (i.e., BERT base, BERTlarge, BERTweetbase, BERTweet-COVID-19, and BERTweet-large) to build a classifier to detect misinformation related to COVID-19 prevention with garlic consumption (Kim et al., 2022). The authors tested these models on two datasets, the “COVID-19 Rumor Dataset” and the “Garlic-Specific Dataset”. The COVID-19 Rumor

Dataset consists of 6834 labeled Tweets and news posts harvested by Chang et al (2021). The Garlic-Specific Dataset consists of 17711 tweets that mentioned COVID-19 and garlic in the same Tweet from November 2019 to April 2020. The Tweets were then further filtered to only include Tweets in the English language. The remaining corpus was then split into 70/30 for labeling. The study utilized two labelers to classify the Tweets as "misinformation" or "other". Name handles (e.g., @name) were removed, tweets were cleaned, and lemmatized.

After labeling the data set was split into 80/20 for testing and training respectively for fine-tuning each BERT model with what appears to be the default parameters on the Huggingface.co repository. However, they trained their models over eight epochs. Results of their fine-tuning versions varied significantly. For the COVID-19 Rumor data, the best performing model (BERTweet-COVID-19) only achieved 0.647 and 0.588 respectively, and F1 scores. BERTweet-large achieved the best performance and achieved an accuracy of 0.911 and an F1 score of 0.894. The differences in the performance of these two data sets are an interesting demonstration of how labeled data and fine-tuning can vary drastically. Though the authors offer thoughtful limitations to their study, it is challenging to pinpoint the direct cause of poor performance while testing the COVID-19 Rumor data test because some important diagnostic parameters (e.g., validation loss and training loss) were not mentioned in the article. Furthermore, the challenges of labeling data are often overlooked by researchers that are eager to test state-of-the-art algorithms. Problems with consistency labeling data occur not only in NLP matters but also in many other domains relying on supervised data. If data are labeled inconsistently, the same inconsistency will more than likely be reflected in a model. Moreover, even if the algorithm may find patterns and understanding in a training set, it does not guarantee that a human researcher would, or vice versa (e.g., sarcasm). That being, the results for the Garlic-Specific Dataset with BERTweet-large were excellent.

Conclusion

This chapter offers an overview of the popular methodologies, employed historically and currently in the field of natural language processing. This section also described in detail how several studies have successfully applied these techniques to studies of public health surveillance, and disease outbreak monitoring, especially related to the COVID-19 pandemic. Truly groundbreaking advancements have occurred in the field of NLP in very recent years and are continually growing and accelerating at an astonishing pace in both industry and academia. Finally, the literature discussed in this chapter demonstrates how researchers quickly adapt and continue to drive innovation to solve problems. It will be exciting to observe and contribute to these technologies that will ultimately help improve and save lives.

CHAPTER III:

METHODOLOGY

- Melton, C. A., Olusanya, O. A., Ammar, N., & Shaban-Nejad, A. (2021). Public sentiment analysis and topic modeling regarding COVID-19 vaccines on the Reddit social media platform: A call to action for strengthening vaccine confidence. Journal of Infection and Public Health, 14(10), 1505-1512.

- Melton, C., Olusanya, O. A., & Shaban-Nejad, A. (2021). Network analysis of COVID-19 vaccine misinformation on social media. *Stud Health Technol Inform*, *287*, 165-166.

- Chad A. Melton, Jintae Bae, Olufunto A. Olusanya, Jon Hael Brena, Eun Kyong Shin, and Arash Shaban-Nejad. "Semantic Network and Content Analysis of COVID-19 Vaccine Related Social Media Text". To appear in Proceedings of the AAAI International Workshop on Health Intelligence (W3PHIAI 2022). Vancouver, Canada. Feb 22- March 1, 2022

- Melton, C.A, White, BM., Davis, RL., Bednarczyk, RA., Shaban-Nejad, A. "Sentiment Analysis of COVID-19 Vaccines: A Comparative Study of Twitter and Reddit Social Media Platforms". Journal of Medical Internet Research 2022 (*in review*).

-

Description of Data

Data Overview

This study compared COVID-19 vaccine-related material from two popular social media platforms, Reddit and Twitter from January 1, 2020, to March 1, 2022. These two platforms were chosen due to their worldwide usage, vibrant discussions, and high user count (approximately 450 million). The Reddit platform is composed of user-created communities (subreddits), in which members adhere to a set of community regulations. Subreddit members have the option to post links, images, videos, and text. Community members then typically "upvote" or "downvote" on a post based on their opinion of the quality of that post and/or leave comments. Depending on the distribution of votes, posts are classified as hot, new, rising, and controversial. The most popular posts within each category are then moved to the top of the community page. These comments are subjected to the same vote ranking system. The upvote/downvote system within Reddit is intended to increase the quality of the posts to minimize non-relevant material. The Reddit user base is expected to be about 60 percent male, in their mid to late twenties, 50 percent in the US, and typically educated. It is estimated that 70 percent of Reddit users are Caucasian, 12 percent Hispanic, and 7 percent are of African descent (Todorov 2022).

Twitter is more individualized than based on distinct communities. Similar to Reddit, Twitter users can post or "Tweet", as well as like, share, retweet, and comment. However, a Tweet is more visible based on the number of followers an author has rather than the popularity of a post (i.e., Reddit). Approximately 17 percent of Twitter users are located in the US, typically younger than 50 years old. It is also estimated that about 42 percent of Twitter users have at least a bachelor's degree and 42 percent are female (Wojcik and Hughs, 2019). The timeframe included the earliest parts of the pandemic to trace the evolution of sentiments over time. Most importantly, these platforms were chosen because only a small number of comparative studies focused on the typical user, especially those related to COVID-19 vaccine sentiment or other vaccine content.

Significant effort was taken to identify and remove Twitter postings that were found to be directly from news agencies or bots. These posts were identified due to an overwhelmingly high post count during the 26 months period relative to the average posting of a "normal" user as well as by visually inspecting tweets of users that appeared at an abnormal frequency. Both Twitter and Reddit (R.2) datasets were limited to only include users who posted less than or equal to 200 times throughout the timeframe. This value was determined after rigorous data exploration and manual inspecting of the datasets. These steps were important due to the repetitive nature of many bot tweets which had the potential to skew sentiment calculations and misalign the goal to compare the normal user base of both platforms. Though methodologies in harvesting Reddit and Twitter data differ slightly, both datasets underwent similar cleaning steps. Both were queried for the same relevant terms typically present in online discussions about COVID-19 vaccines. This step was important due to the tendency for some extended comment threads to meander off-topic. This occurrence was especially true with threads from some Reddit communities. The posting frequency of the two platforms was relatively similar in the early months of the pandemic.

Frequency increased dramatically for both platforms in late September 2020-October 2020 as news of the vaccine roll became more widespread. Though each platform displays four spikes in posting frequency a similar times (Oct 2020, Mar-Apr 2021, Aug-Sep 2021, Dec 2021-Jan 2022) they each obtain a maximum in different months. Reddit reached its maximum posting in Mar-April 2021 while Twitter reached its maximum in Sep-Oct 2021.

Twitter Data

Approximately 13 million Tweets were harvested from January 1, 2020 to March 1, 2022 using the snscrape and Tweepy API Python libraries. After removing tweets by suspected bots, news media, highly repetitive-high frequency users, or duplicate Tweets, the final Twitter data set consisted of 9,518,270 Tweets authored by 3,006,075 Twitter users. The Tweets contained a total of approximately 16.32 million likes with a maximum of 430,758 and an average of 14.9. Tweets cannot be downvoted but approximately 4,794,865 had zero likes attributed. Statistics on tweet sharing or retweets were not collected because this metric was not available for both platforms (see Fig 1).

Reddit Data

Reddit data were gathered in three phases of this study and therefore there are three Reddit (Reddit 1.0, Reddit 1.2, and Reddit 2.0) datasets. Multiple data harvests were conducted because this research was conducted in the midst of the pandemic. For the Reddit 1.0 dataset, I harvested approximately 18,000 posts from thirteen subreddits through the Reddit API on May 16, 2021. Because Reddit communities potentially contain some inherent bias due to strict community rules, as well as content monitoring by a moderator, these subreddits were chosen to create a non-biased dataset from a diverse selection of communities that vary widely in political views as well as positions on vaccination. These subreddits were also chosen due to the large number of members (approximately five million members). Data were cleaned first by combining each subreddit into a centralized database.

The data were then organized by date and then queried for terms specifically related to the COVID-19 vaccine. These terms included COVID vaccine, vaccine, vaccination, immune, immunity, COVID vaccination, corona vaccine, COVID19 vaccination, COVID-19 vaccination, coronavirus vaccination, coronavirus vaccine, COVID-19 vaccine, coronavirus vaccine, coronavirus vaccination, Moderna, Pfizer, J&J, Johnson & Johnson, COVID vax, corona vax, covid-19 vax, covid19 vax, coronavirus vax). The finalized dataset consisted of 1401 posts and 10,240 comments (11,641 in total) written by greater than or equal to 8281 authors/users, 1048 of whom posted multiple times. In actuality, the number of authors could have been as high as 9013. These additional users are probable because Reddit removes the user ID from posts after a user deletes their account. However, the post content and upvotes remain. After data were cleaned and organized, I conducted a sentiment analysis and LDA topic modeling with other NLP tools in Python.

For Reddit 1.2, I harvested approximately 300,000 posts and comments from 12 online *subreddits* on the social media platform on October 31, 2021. Reddit 1.2 is slightly different in the time frame but does not include r/NoNewNormal and r/coronavaccine. r/NoNewNormal was removed by Reddit due to threatening behavior demonstrated by many community members. r/coronavaccine merged with r/CovidVaccinated which is still included in this subset. Similar to Reddit 1.0, data were cleaned and queried for posts related to COVID-19 vaccines/vaccination. The final dataset consisted of 31432 posts/comments authored by 20429 users between January 1, 2020, and Oct 31, 2021. Finally, these posts/comments received a total of approximately 1.26 million votes, indicating a high degree of community interaction. The majority of subreddits were similar in posting frequency for the first months of the timeframe. Posting in several subreddits rapidly increased over time (see Fig 2). For the Reddit 2.0 dataset, I harvested 579,241 user-created posts from 67 subreddits (see Appendix for the list of Subreddits) with the Python Reddit API Wrapper(PRAW) on March 2, 2022. These subreddits were collected to gain a broad understanding of sentiments related to the COVID-19 vaccines as well as to avoid potential biases in data collection. These subreddits contained a total of 5,590,913 subscribers as of March 1, 2022. The query process removed a large portion of terms not related to the terms. After visually inspecting and confirming the results of the querying process, the final Reddit data set consisted of 69,079 comments composed by at least 9,932 authors. These posts contained a total number of approximately 2.2 million upvotes with

an average of approximately 31 upvotes and a maximum of 18,253. After cleaning the datasets, both were processed with each fine-tuned model. The models reported polarity (i.e., positive and negative) and a confidence score (-1 for most negative and 1 for most positive) for each of the entries (see Fig 3, and Table 1).

Analytical Methods

Sentiment Analysis for Reddit 1.0

This portion of the study used a lexical-based sentiment analysis. This method employs dictionaries of words with a previously assigned valence score as a reference for the text analysis. This design is somewhat similar to using labeled historical data in machine learning but computes much faster because dictionaries have been pre-trained. In general, sentiment can be determined from several levels of complexity ranging from large volumes of text to single words or unigram. First, the Regex library was employed to clean and remove special characters or any remaining hyperlinks in the text of each subset. At this point, subjectivity and polarity were calculated with the TextBlob subjectivity and polarity functions. The subjectivity function returns a floating-point value between [0,1](0 being the most factual and 1 being the most opinionated). The function works by quantifying modifiers or adverbs in a sentence *(e.g., extremely lethargic*). Subjectivity values measuring between (0.4,0.6) were classified as neutral, values greater than 0.6 were classified as ¨Highly Opinionated¨, and less than 0.4 were classified as "Least Opinionated". Polarity returns a floating-point value between [−1.0,1.0] where −1.0 is considered to be the most negative while 1.0 is the most positive. The polarity tool works by comparing each word in a user-provided corpus to a previously defined polarity reference dictionary within the TextBlob.sentiment.polarity constructs.

Sentiment Analysis for Reddit 2.0 and Twitter Data

For this study, I chose to explore the capabilities of DistilRoberta. RoBERTa is a more robust model than BERT and DistilRoberta is an optimized version of RoBERTa.[18,19] Developed at Facebook, RoBERTa was trained on 160 GB of text compared to 16 GB of BERT. RoBERTa dropped the next sentence prediction feature of BERT and added dynamic token masking during training. These enhancements are estimated to have improved OB performance significantly (2-20 percent) (Sanh et al., 2019).[19] Compared to RoBERTa, DistilRoBERTa was trained on approximately 40 GB of text data (OpenWebTextCorpus) and operates about twice as fast but loses 3 percent of BERT's performance.

Since time is of the essence in a global pandemic, combined with the fact that labeling data is time-consuming and costly, I created a custom training data set by labeling sentiment (positive or negative) for approximately 3600 tweets related to COVID-19 vaccines. I chose to label tweets exclusively for this study, because the 280-character limit of a tweet (i.e., compared to a Reddit post limit of a maximum of 10,000 characters) would allow our small team to create a time-relevant training data set more quickly. I then augmented our data set through the process of back-translation with several language models on the Hugging Face model repository. Back-translation was chosen after testing a few other methods of text augmentation. Some techniques (e.g., word masking) resulted in far more duplicated texts that would eventually need to be removed. Back-translation relies on subtle differences between language structure, word meaning, and syntax. In effect, the outputted text will vary slightly from the inputted text without losing semantic and contextual meaning (Beddiar et al., 2021). In our case, the back-translation method translated our English-language text into a language (e.g., French, Chinese, Greek, and Hebrew) and then back into English. After removing duplicates, our final augmented data set consisted of 48,691 tweets.

I fine-tuned DistilRoBERTa-base via the Huggingface *Trainer* class which provides the user with an API for training with *PyTorch.* The data were then randomized and segregated into 40k training tweets, 4k validation tweets, and 4,691 for testing. Training hyperparameters included a 2e-05 learning rate, 32 training and evaluation batch size, a seed number of 42, and a linear scheduler with 500 warmup steps. I used the Adam optimizer with betas [0.9m 0.999], and epsilon of 1e-08. Lastly, the model was trained for three epochs. These hyperparameters achieved a training loss of 0.1284 validation loss of 0.1167, a precision of 0.9561, an f1 score of 0.9592, and an accuracy of 0.9592 (see Table 2).

Following the fine-tuning of the model, I processed the Twitter and Reddit data through the Huggingface *pipeline* for sentiment analysis. The model returned a label of either *positive* or *negative* for each Tweet or Reddit comment. Along with the determined polarity, the model also returned a probabilistic confidence score ranging from [0,1]. For clarity, Tweets or comments classified as negative were multiplied by -1 to reflect the negative sentiment (see Fig 4).

Topic Modeling with LDA

The Gensim LDAModel algorithm was used to create (Latent Dirichlet Allocation) LDA models for each month (Srinivasa-Desikan, 2018., Gensim, 2011., Blei et al., 2003). This technique is highly useful in detecting latent topics in large textual data. LDA assumes that documents with similar topics will use similar diction and that the topics will display a sparse Dirichlet distribution. For example, every word in the document is randomly assigned to a user-defined number of topics T. The algorithm then calculates the proportion of words in each document assigned to a topic (i.e., [p(topic T | document D)]) and then the proportion of words that were assigned to a topic overall document (i.e., [p(word W | topic T)]). The product of these pro-portions is computed for each topic T and compared to every other topic T until algorithmic convergence is achieved (Blei et al., 2003). After removing stop words (e.g., determiners, conjunctions, and prepositions), and lemmatizing the corpus (i.e., converting a word to its base form), coherence values were tested on 50 different LDA models to determine the most statically appropriate number of probable latent topics. Though coherence values are insightful, topics were qualitatively analyzed to double-check for content coherency rather than numerical. Because the data set was collected from posts ranging over approximately six months, I conducted the sentiment analysis and LDA topic modeling using collective data ranging from December 1st, 2020 to May 15, 2021, as well as individual months. Once polarity was determined, data were divided by polarity (i.e., positive, negative, neutral) and conducted further LDA based on the previously calculated polarity (see Fig 5).

Semantic Network Analysis

Network analysis can unveil the network structure and reveal vaccine hesitancy discourse in online discussions. To accomplish this task, I conducted a temporal semantic network analysis to observe graph evolution over time. Network analysis uses a graphical representation of nodes and edges to provide insight into data that may not be observable upon the surface. In computational semantic network analysis, *nodes* represent *tokenized* words while *Edges* represent a connectedness between nodes. A multistep analysis was conducted with the *Python* library, *Networkx* (Networkx, 2012) (Srinivasa-Desikan, B. (2018)).

Due to the large dataset of over 9.5 million Tweets, previously constructed Python code was converted to be utilized in a *high performance computing* (HPC) environment. In this case, I used the ISAAC Next Generation Cluster at the University of Tennessee, Knoxville. HPC used a *divide and conquer* approach to computation where data and calculations are distributed amongst a certain number of nodes and cores. My code was written so that the Twitter data were to be processed 96 cores, and the Reddit 2.0 data were processed over 48 cores. For example, October 2020 from the Reddit 2.0 dataset contained about 1500 comments. On a single Intel Core i7-8750 @ 2.20 GHz, this month of comments took around 680 seconds to run through all the steps needed for this analysis. Thanks to these 48 to 96 cores from of ISAAC, the entire network building steps were speed up between 4 and 5 times respectively (Amdahl 1967) .

After removing *stop words* (e.g., determiners, conjunctions, and prepositions*)*, lemmatizing the corpus (i.e., converting a word to its base form), and vectorization, networks were created for each month in Reddit 2.0 and the Twitter data sets. Lastly *betweenness centrality* (BC) and *eigenvector centrality* (EC) were calculated for each node. Essentially, BC displays the importance of a node (i.e., word) based on calculating the number of times a node is included in the shortest route between other nodes and EC measures node influence based on quantity of connections to other nodes (ideal for social networks).

(Linton 1977). Due to the *hairball* effect that occurred while including each word in a monthly corpus, networks included in this document were limited to less than or equal to 160 nodes by setting an appropriate weight threshold. To detect the tendency for clustering to occur within the data, transitivity was calculated for the complete corpus as well as for the scaled-down versions with limited node counts. Lastly, graph density and networks/subreddit statistics were collected and networks were also visually inspected to verify coherence in the interpretation of the results (see Fig 6).

CHAPTER IV:

ANALYTICS & EVALUATION

- Melton, C. A., Olusanya, O. A., Ammar, N., & Shaban-Nejad, A. (2021). Public sentiment analysis and topic modeling regarding COVID-19 vaccines on the Reddit social media platform: A call to action for strengthening vaccine confidence. Journal of Infection and Public Health, 14(10), 1505-1512.

- Melton, C., Olusanya, O. A., & Shaban-Nejad, A. (2021). Network analysis of COVID-19 vaccine misinformation on social media. *Stud Health Technol Inform*, *287*, 165-166.

- Chad A. Melton, Jintae Bae, Olufunto A. Olusanya, Jon Hael Brena, Eun Kyong Shin, and Arash Shaban-Nejad. "Semantic Network and Content Analysis of COVID-19 Vaccine Related Social Media Text". To appear in Proceedings of the AAAI International Workshop on Health Intelligence (W3PHIAI 2022). Vancouver, Canada. Feb 22- March 1, 2022

- Melton, C.A, White, BM., Davis, RL., Bednarczyk, RA., Shaban-Nejad, A. "Sentiment Analysis of COVID-19 Vaccines: A Comparative Study of Twitter and Reddit Social Media Platforms". Journal of Medical Internet Research 2022 (*in review*).

Overview

This chapter offers an in-depth presentation of the results of each step of this research. The first section specifically discusses the results of the *Textblob* sentiment analysis, subjectivity analysis, and LDA topic modeling of the Reddit 1.0 data set followed by a discussion of a semantic network analysis of the Reddit 1.2 data set. The following two sections discusses the results of my fine-tuned DistilRoBERTa sentiment classifier, along with the LDA analysis and semantic network analysis first for the Reddit 2.0 dataset (Reddit), and then followed by the Twitter Coronavirus Data set.

Results

Reddit 1.0 Sentiment Analysis

For the combined dataset, the polarity analysis found that 56.68% of the posts measured positive, 27.69% were negative, and 15.63% neutral. The mean polarity value reported was 0.0520 and the variance was 0.0415. The subjectivity analysis reported that 73.15% of the comments measured in between [0.25, 0.75] and considered "neutrally subjective", 18.13% were reported to be minimally subjective (less than 0.25) while the remaining 8.72% were highly subjective (greater than 0.75). The mean subjectivity and variance were reported to be 0.4450 and 0.0560 respectively. The comments from these posts received a total of 612,217 upvotes. Upvote scoring ranged from a maximum value of 11,110 upvotes and a minimum of −135. The mean upvote count was reported at 50.04, and the mode was 7 upvotes. Comments that were classified as negative received 133,305 upvotes (23.24%). The Neutral classified comments received a total of 94,641 upvotes (16.50%). Lastly, positive classified comments received 345,607 upvotes (60.26%) (see Fig 7, and Fig 8).

Reddit 1.0 Topic Modeling

The combined dataset LDA modeling returned an optimal number of five latent topics in the dataset. This classification was determined by computing coherence scores for 50 models. The calculation returned a high score of 0.5641 (see Fig 9, and Fig 10).

For the collective dataset, the LDA modeling results displayed a total of five optimal latent topics. *Topics 1-4* are closely related to a broader discussion of the vaccine, safety concerns, efficacy, and potential side effects. Side effects mentioned in these topics ranged from being less severe (e.g., fever, sore) to death. Keywords in these topics could suggest that these specific users consist of people that have not taken the vaccine at the time of content composition and are expressing their concerns about taking the vaccine or are directly related to the discussion of side effects experienced by users who have received at least one dose of the vaccine. *Topic 5* seems to be focused on much broader terms, information (i.e., *news, source, question*) as well as a direct mention of concerns about vaccination. The topic also mentioned *autism*, most likely in reference to the antivaccine movement's fixation on the false narrative that vaccines cause autism. The findings in *Topic 5* are particularly interesting due to the direct detection of discussions related to vaccine misinformation and autism. For example, a positive-polarity comment sarcastically stated "*Vaccines cause autism, huh? Well, I've got two more days till I get completely upgraded (I get the second Moderna dose on Tuesday)*". While others believed the misinformation such as the content referenced in this negative-polarity post *"MRNA is not safe and should be put into your body unless you are* 10000% *what it is. mRNA can do/mean literally anything from protecting you from COVID to sterilization all the way into making you autistic (not like the anti-vaxxers thinking the influenza vaccine is the cause for autism-like MRNA can literally be edited to give you mental problems)".* These topics generally focused on questions about the vaccine, side effects, medication, experiences with the vaccines, and intervals of time.

The sentiment analysis results for individual months agreed with the combined analysis and reported that the majority of posts were positive. December reported 57.63% positive, 26.21% negative, and 16.16% neutral. January reported 59.49% positive, 25.52%

negative, and 14.99% neutral. February reported 57.93% positive, 28.10% negative, and 13.97% neutral. March reported 57.13% positive, 25.97% negative, and 16.9% neutral. April reported 54.98% positive, 28.96% negative, and 16.06% neutral. Lastly, May reported the least positive sentiment and highest negative sentiment at 53.05% positive, 30.57% negative, and 16.38% neutral.

Overall, the LDA topic modeling results were similar to the complete data set. However, in this case, latent topic quantities were much smaller due to the smaller corpus with each month (less than or equal to three latent topics). The content of these individual months was very similar to the overall combined data set except for December. Latent topics in December contain keywords associated with vaccine trials, group, efficacy, as well as potential side effects. January and February both display latent topics related to vaccine dosage, the number of doses, immunity, and side effects. March, April, and May are more closely related. Though these months include topics similarly detected to December through February, these months reference latent topics more directly related to hesitancy (i.e., concern, risk), as well as death. Interestingly, some discussion of T cells was detected in April and May (see Table 3).

The sentiment topic modeling results were significantly more convoluted than the combined and monthly topic model. The models in these three polarities all contain some themes in common related to discussions and questions about the vaccination process, side effects, concerns, time, and immunity. The negatively classified post topics contained additional keywords such as *government, state, science*, *employee*, *risks,* and *several expletives*. Posts classified as neutral displayed topics that referenced *physicians, Pfizer,* *research*, *video*, *link*, *issue*, and *stories*. Lastly, topics related to the positive posts included terms such as *Moderna*, *flood*, *safe*, *woman*, *pregnant*, *family*, and *response*. The positive posts also contained keywords related to death as well as expletives (see Table 4).

Reddit 1.2 Semantic Network Analysis

As expected, changes in the networks reflect the dynamic conditions and events that have occurred since the first COVID-19 cases were detected. Semantic and network structural changes are observable in the *Giant* and *VE* networks in several significant shifts as different phases of the pandemic came and went. For example, in early Jan 2020, a small number of nodes are visible representing words associated with COVID-19 but the *Giant* network does not display interconnectedness between the node *Vaccine* and other nodes representing COVID-19. These occurrences rapidly increase as infection rates climb and online discussion shifts towards vaccines for COVID-19. Nodes tend to reflect conversation regarding side effects (e.g., fever, sore arm, body aches, etc.) as vaccines become more readily available. In the dataset, the large increase in posts from r/CovidVaccinated in April 2021 contributes to vaccine side-effects interconnectedness and appearance as well. Moreover, this occurrence was also expected based on previous topic modeling studies (Melton et al., 2021). Unfortunately, nodes representing misinformation keywords become more apparent as the interconnectedness with the node "Vaccines" increases in conjunction with COVID-19 keywords (e.g., Vitamin D, autism, Bill Gates, Big Pharma) over several months. Visual analysis of the raw text data suggested a wide range of vaccine hesitancy behaviors as well. These behaviors included hesitancy due to fear of vaccine side effects, feelings of "threatened freedoms", false expertise, ignorance of how vaccines work, “big pharma-motivated pandemic conspiracy, anti-vaccination beliefs, and many others.

Overall, the *Giant* network node and edge count generally increased with time, exhibiting fluctuations in certain months. This increase occurred simply due to more frequent postings of time. The *Vaccine Ego (VE)* network behaved similarly although the node and edge count were much less due to the nature of an ego network. Transitivity values for the complete corpus or *Giant* network decreased over time and ranged between 0.13 (Dec 2020) and 0.21 (Jan 2020) with an average value of 0.15. Greater node clustering occurred within this downtrend during Aug 2020, Oct 2020, April 2021, and July 2021. The *VE* net exhibited similar characteristics with increased clustering in June 2020, Oct 2020, April 2021, and July 2021. Density for both *Giant* and *VE* networks tended to ebb and flow from month to month but decreased over time as well due to an increasing post quantity (see Fig 11).

For the betweenness centrality networks, centrality ranged from 0.21 for the node *immunity* to 0.892 for the node *vaccine*. Variations of the word “vaccine” (i.e., vaccine, vaccines) occurred in the top 10 highest BC values throughout each month in the dataset. Furthermore, I observed changes in centrality values related to nodes that represent some terms common in COVID-19 misinformation. For example, in September 2020 (see Fig 4), the nodes *vitamin* and *d* are connected to a few terms related to COVID19. Centrality for the nodes was calculated to be 0.16 and 0.20 respectively and amongst the top five for the month. As the occurrence of these two nodes diffuses over time, the centrality values diminished substantially to *vitamin* (0.0004725 and rank 513/5894) and *d* (0.00055 and rank 461/5894) in October 2020. However, *d* (0.001193 and rank 253/9301) and *vitamin* (0.00168 and rank 177/9301) both increased but increase again in December 2020. Nodes indicating vaccine hesitancy were also observable in the networks and represented by example keywords such as *scared* or *worried*. However, visual inspection of some comments revealed the intent of vaccination even though the user experienced anxiety related to the vaccine. For example, a comment from April 2021 said, "I am going to get my shot today! Half excited, half scared. Not scared from like conspiracy theory stuff lol, but I have had systemic allergic reactions before, so yeah a little nervous there."(see Fig 12).

Reddit 2.0 Sentiment Analysis

The Reddit sentiment polarity analysis for the DistilRoBERTa fine-tuned polarity analysis found approximately 37 percent (25780) of posts to be classified as negative and 63 percent (42473) to be classified as positive. The highest polarity reported in the experiment, the maximum positive rating occurred in April 2021 (approximately 73.1 percent) and the minimum positive rating occurred recently in February 2020 (48.4 percent). For the confidence scores, the positive classified comments had a maximum score of 0.999, a minimum of approximately 0 (1.55e -4), and a mean of 0.870. The negative classified scores rated the most negative comment at -0.999, the maximum value of approximately 0, and the mean of -0.808. (see Fig 13, Fig 14, Fig 15).

Reddit 2.0 Topic Modeling

Latent topics detected with LDA for the Reddit 2.0 data analysis changed as the pandemic progressed and vaccines were developed. Conversations ranging from January

2020 to March 2020 included discussions about vaccination but related to influenza. Though “coronavirus” was being discussed early on, significant mentions of the virus were infrequent until March 2020. Furthermore, significant discussions of COVID-19 vaccines were not prevalent until June 2020. As vaccines were developed, latent topics naming vaccine manufacturers were detected in November 2020. In January 2021, significant discussions of the vaccine side effects begin to dominate the majority of topics as the vaccine rollout occurred. Other discussions involving the variants and deaths increased after January 2021 as well. Topics related to boosters were detected first in Oct 2021 and persisted through the remaining months in the data. Interestingly, the majority of months contained somewhat homogenous latent topics across positive and negative classified sentiment.

Reddit 2.0 Semantic Network Analysis

The networks created from the Reddit 2.0 dataset display a somewhat decentralized structure for each month. Beginning in January 2020, the smallest network consisted of four nodes (i.e., flu, shot, vaccinated, and people) that represent discussions regarding the flu vaccine. The network for February 2020 expands significantly. The first mention of coronavirus appears this month and nodes representing flu are still present as well. As the virus migrates to populations across Asia, Europe, Africa, and the Americas, networks for March – Jun 2020 contain nodes representing terms related to coronavirus infections, treatments, symptoms, quarantine, social distancing, and death. This time also contains the first appearance of alternative treatments and prevention (i.e., vitamin →d in April 2020). Moving into the summer months of 2020, the network structure expands to reveal more interesting node relationships such as herd and immunity, cases and death, and others with continuing mention of alternative therapies such as ivermectin and Vitamin d. During these months, the node representing Moderna is observable for the first time (June 2020). In the fall of 2020, nodes symbolic of vaccine trials (e.g., clinical, trial, placebo, participants) begin to cluster significantly. As news of the vaccine trial results is released, the node Vaccine begins to take precedence over other nodes and eventually forms a single supercluster for November and December of 2020. Moreover, this month potentially reveals an indication of nodal relationships representing *long covid*. The structure appears to undulate from January 2021 into the spring and summer months. However, nodes representing vaccination, and vaccine side-effects tend to dominate in frequency.

This timeframe also sees continued mentioning of vaccine manufacturers, as social medial users consulted each other on which vaccine they would take. The occurrence of nodes representing alternative medicine also continues to be less frequent during this time frame. July 2020 contains nodes representing similar words as the previous post-vaccine release months. During this time, the nodes *vaccinated* and *unvaccinated* occur relatively close together. As the school year begins, nodes representing anti-masking are apparent, although away from the central cluster.

Late 2021 continues to contain nodal relationships similar to the summer and late spring but the network structure changes somewhat due to the appearance of nodes representing geographic locations. This occurrence is somewhat expected because many subreddits included centered around specific cities or states were included in the dataset. The network structure continues to decentralize in January and February 2022 and appears somewhat similar to earlier months in the pandemic. Furthermore, nodal relationships indicative of mental health discussions are apparent in some clusters. These discussions are mainly based on high betweenness centrality nodes (i.e., health and coping). These nodes are connected to other smaller nodes such as mental, stress, managing, and life (see Fig 28-32).

For the top five BC values, the results display high centrality values for nodes related to the flu and flu vaccine in the first for February and March 2020. The first mention of the coronavirus was detected in April 2020 and typically was typically observable through October 2020. Leading up to the late fall, nodes representing the word cases, immunity, vaccines, and death are also present during the summer to early fall of 2020. The top five centrality values shift to nodes representing vaccines in November 2020, and migrate to discussions related to vaccine doses, and side effects as the spring of 2021 progresses into the summer and fall months of 2021. Though the node *vaccine* is present through the remaining months of the dataset, the top five betweenness centrality nodes tend to represent words such as cases, health, people, and coping. It is noteworthy to mention that the node vaccine is present in the top five for each month of the data set except for February 2020 and September 2020 (see Table 5). The results of the EC nodes were similar to the top BC nodes and only differed by approximately 10 words. Though some node values were almost identical, EC values were often lower. For example, in January 2022, the node “cases” had an EC value of 0.27, while the BC value for the same node was 0.32. For the BC values the Reddit 2.0 dataset, the minimum value was 0.036, a maximum was

0.98 , the average was 0.21, and the variance was 0.03. The EC displayed a minimum value of 0.45, a maximum of 0.69 , an average of 0.23, and a variance of 0.02.

Transitivity for the Reddit 2.0 dataset begins modestly at 0.15 with minimal clustering. A peak with a short retrace quickly begins to rise as news of the virus and infections spread until June 2020 when the transitivity reaches its highest point (~0.39). This peak then quickly falls over the next few months and reaches global minimum transitivity (~0.05) in September 2020. A sharp increase occurs as the next peak (~0.20) occurs in Oct 2020, followed by a modest decline. As vaccines are released to the public, the transitivity significantly climbs until the next two peaks occur in April 2021 (~0.30), and June (~0.34). Transitivity levels stay between (0.22 – 0.31) for the remainder of the time frame. Though slight fluctuation in transitivity could be a result of density changes, it is more likely the fluctuations are a result of evolving topics within the datasets throughout the pandemic.

Twitter Sentiment Analysis

The DistilRoBERTa fine-tuned polarity analysis was determined to be more negative (approximately 55 percent, 5215830 Tweets) than positive (approximately 45 percent, 4302440) throughout the timeframe (see Fig 16). The maximum positive rating occurred in March 2021 (approximately 55 percent). However, the minimum polarity occurred in January of 2022 (approximately 64 percent), displaying a steady decrease in polarity from the maximum. For the confidence score, the positive classified Tweets had a maximum score of 0.999, a minimum of approximately 0 (3.58e-7), and a mean of 0.868.

The negative classified scores rated the most negative Tweet at -0.999, the maximum value of approximately zero (-1.78e-6), and the mean of -0.882 (See Fig 17 and Fig 18).

Twitter Topic Modeling

As the pandemic progressed, the LDA models for the Twitter data widely reflect similar changes in the evolution of latent topics as the Reddit 2.0 dataset. Beginning in January 2020, topics discussed in positive and negative classified Tweets mainly focus on broad discussions of immunology and immunotherapy with no mention of COVID-19 specifically. The focus of conversations quickly changes in February 2020 to discussions about COVID-19, treatments, vaccination, and progress until discussions begin to shift focus when news of vaccine development is released. As the vaccines were distributed to the general public in March 2021, discussions regarding eligibility, side effects, safety, and vaccine manufacturer were observable. Some mention of serious side effects were detected as well. For example, discussions regarding blood clots associated with the Jansen vaccine were detected in April 2021, as six women developed the rare condition (Ledford 2021). As the death toll dramatically increased, discussion of COVID-19-related deaths become prevalent in most months after May 2020. Furthermore, topics were detected throughout this data set related to major events in specific geographic locations. Lastly, only minor differences in latent topics were detected between the positive classified and negative classified Tweets.

Twitter Network Analysis

Similar to the Reddit 1.0 data set, the *Giant* network node and edge count increased with time and exhibited some fluctuations. Because January 2020 contained minimal mention of any terms related to COVID-19, Coronavirus, or associated vaccines, meaningful networks could not be created with the earliest month in our data set. However, this frequency quickly changes in February 2020 as the COVID-19 infections spread rapidly. Strong interconnectedness (ie., high degree centrality) was observed between terms related to COVID-19, and vaccines/vaccination for every month in the dataset. As time progresses into the fall of 2020, edges extend from these central structures to nodes representing terms related to specific geographic locations (e.g., India, Canada, China), vaccine development news and distribution, long-term disease effects (pre-long COVID), immunity, and deaths (November 2020). As certain populations and politicians expressed anti-vaccine and/or vaccine-hesitant sentiment, nodes related to mandates and alternative COVID-19 treatments are present (e.g., covid → ivermectin) (see Aug 2021). Interestingly, the node representing "flu" is present every month. However, the centrality is significantly and consistently lower (less than ~0.24) for the entirety of the dataset (see com/Cheltone/Twitter_Reddit_C19_Networks, and Fig 23-27).

For the betweenness centrality analysis, the terms *COVID-19* and *vaccine* remain in the top five highest betweenness centrality values through the time in focus. Similar to the LDA results, higher centrality values were detected in association with the release of vaccine development news (e.g., Moderna press release Nov 17, 2020) (Mahase 2020). This analysis was also successful at detecting discussions regarding COVID-19 infections of two politicians including the former United States president (2016-2020) and the Canadian prime minister. Similar to the Reddit 2.0 EC analysis, BC and EC node occurrence was similar and only differed by approximately 15 words. Again, some node values were almost identical but EC values were still often lower. For example, in February 2021, the node “vaccines” had an EC value of 0.50, while the BC value for the same node was 0.73. The BC values of the Twitter data were a minimum value was 0, a maximum was 0.93 , the average was 0.22, and the variance was 0.07. The EC displayed a minimum value of 0.10, a maximum of 0.77 , an average of 0.31, and a variance of 0.02.

Transitivity analysis of the Twitter *Giant Network* suggested that the topics of discussion oscillated throughout the pandemic but displayed four significant increases (not including the initial high transitivity). The first peak begins to form in March 2020 and reaches the maximum of 0.25 in June 2020. A steep decline is observable immediately after, almost reaching initial levels in July 2020. Transitivity remains between approximately 0.05 - 0.10 until the second peak begins to rise in Dec 2020 (initial vaccine release) (FDA 2021) until the third major increase in March 2021 (general public release) when transitivity increases to 0.11. Clustering dissolves slightly moving into the summer but begins to increase again with the increase of Omicron infections and peaks at 0.10 in November 2021. Lastly, transitivity begins to increase further in January 2022 and reached a level of 0.117 at the end of the dataset (see Figure 19 and Table 6).

Covid-19 Vaccine Sentiment Expressed on Reddit and Twitter

Overall, the average sentiment for the two social media platforms was similar (52% Reddit vs 53% Twitter). However, insights become observable when looking closely at the month-to-month results. Though sentiment on both platforms oscillated in the early months of the pandemic, Reddit sentiment was lower (ranging from 35-45% positive) from January 2020 until August 2020. Twitter sentiment began higher than Reddit but gradually declined until becoming significantly more negative in October 2020. As Twitter sentiment continued to decline to approximately 35 percent positive, Reddit began a steep increase in polarity in September 2020 and continued to reach the maximum positive sentiment (~57%) in April 2020. Twitter sentiment also began a steep reversal in sentiment from November 2021 until the maximum positive sentiment (~57%) was reached in March 2021. After maximum positive sentiment was reached, both platforms began an oscillating and gradual decline in sentiment to near early pandemic levels (see Fig 20).

It is mentionable to consider the posting frequency of both communities throughout the pandemic when interpreting these results. Very few Tweets mentioned a vaccine for COVID-19 in the first few months of the pandemic, while some subreddits (r/ChinaFlu, Note: the subreddit was founded when the virus was confined to China) had lively discussions early on. Both platforms had a s increase in posting frequency leading up to March - April 2021. However, the frequency of Reddit posting drastically declined while Tweet frequency remained high until reaching its peak in August 2021. This occurrence could be related to the sharing and posting of breaking news events regarding further developments in the pandemic (e.g., boosters, variants).

CHAPTER V: DISCUSSION AND CONCLUSION

- Melton, C. A., Olusanya, O. A., Ammar, N., & Shaban-Nejad, A. (2021). Public sentiment analysis and topic modeling regarding COVID-19 vaccines on the Reddit social media platform: A call to action for strengthening vaccine confidence. Journal of Infection and Public Health, 14(10), 1505-1512.

- Melton, C., Olusanya, O. A., & Shaban-Nejad, A. (2021). Network analysis of COVID-19 vaccine misinformation on social media. *Stud Health Technol Inform*, *287*, 165-166.

- Chad A. Melton, Jintae Bae, Olufunto A. Olusanya, Jon Hael Brena, Eun Kyong Shin, and Arash Shaban-Nejad. "Semantic Network and Content Analysis of COVID-19 Vaccine Related Social Media Text". To appear in Proceedings of the AAAI International Workshop on Health Intelligence (W3PHIAI 2022). Vancouver, Canada. Feb 22- March 1, 2022

- Melton, C.A, White, BM., Davis, RL., Bednarczyk, RA., Shaban-Nejad, A. "Sentiment Analysis of COVID-19 Vaccines: A Comparative Study of Twitter and Reddit Social Media Platforms". Journal of Medical Internet Research 2022 (*in review*).

  Olusanya OA, White B, Melton CA, Shaban-Nejad A. Examining the Implementation of Digital Health to Strengthen the COVID-19 Pandemic Response and Recovery and Scale up Equitable Vaccine Access in African Countries. JMIR Form Res. 2022 May 17;6(5):e34363. doi: 10.2196/34363. PMID: 35512271; PMCID: PMC9116456.

- Olusanya OA, White B, Amuchi B, Melton CA, and Shaban-Nejad A. Perceptions and Misinformation on COVID-19 Mask Mandate During Tennessee School Board Meetings: Analyses and Recommendations for In-Person Learning. Submitted to Journal of *Qualitative Health Research*. **(Under Review)**

Research Outcomes

Interpretation of the Reddit 1.0 Analysis

Although the results displayed in this document suggest that public sentiment in Reddit communities is overall positive regarding discussions about the Covid-19 vaccine or experiences with taking the vaccine, keywords and topics were detected that indicate some hesitancy amongst these users. The results report a higher positive polarity in general, but they do not suggest that the sentiment of these community members has changed significantly during the time interval in focus. This occurrence could be due to the potential bias in these communities and/or related to strict Reddit community guidelines that result in the removal of certain posts, creating either an evidence-based or nonevidence-based echo chamber. It is conceivable that bias could be lessened by amalgamating comments from a right-leaning, left-leaning, and neutral news organization from multiple social media platforms simultaneously (Aiello et al., 2020). Moreover, it is possible that the sentiment analysis reflected the nature of the interaction between users rather than actual feelings about vaccination.

Qualitative analysis revealed the detection of some comments that expressed a negative sentiment toward the vaccine but were given a positive polarity due to certain aspects in the text. For example, the comment, "*Looking forward to being treated like the plague for refusing the – gene-therapy – vaccine. As a proud introvert, I can't wait for people to avoid me!*", received a polarity score of 1. Lastly, it is possible that the results of this *TextBlob* sentiment analysis did not capture some of the nuances in language that the fine-tuned DistilRoBERTa model captured. That being said, it would behoove future researchers to incorporate a multimodal methodology when using TextBlob. Though the lexical based valence dictionary of TextBlob can be utilized more quickly than fine-tuning a model, there is potential for accuracy to suffer. Nonetheless, these results shed light on user activity within these subreddits and suggest that most active community members participate mainly through the upvote/downvote feature. This behavior is demonstrated by the large discrepancy in authors (~9000) compared to comment upvotes (612,217), not to mention the other 4.9 million community members who mainly consume the content without interacting.

Topic modeling quality is often challenging to evaluate because using coherency and perplexity are based on purely numerical relationships in word occurrences. At times, an optimal coherence value may result in topics that are not qualitatively coherent (Chang et al., 2009). Due to this fact, it is fundamentally necessary to inspect returned topics as well as data content. In this study, qualitative analysis more or less agreed with coherence values. LDA results presented in these models kept a common theme over time when considering the month-to-month analysis. Moreover, slight changes in portions of topics are still observable that reflect an evolution in discussion from the early vaccine rollout to vaccines being commonly available. Interestingly, one constant topic that was detected throughout each month, regardless of polarity is side effects. This finding was expected considering many recently vaccinated people discuss and compare side effects on social media as well as in person. Due to the severity of some documented side effects and their wide media coverage, it's highly conceivable that side effects are a major contributor to hesitancy. It is also mentionable that the majority of conspiracy theories were not detectable by the LDA models, indicating a minimal occurrence. Besides the mention of *autism* most manually read conspiracy theories that were sarcastic (e.g., My nanobots must not be working cause my 5G sucks).

Comparing Reddit(Reddit 2.0) and Twitter

This section compares the average Twitter user to a Reddit user and an interesting story of the pandemic on Reddit and Twitter appears with these results. Though the average sentiment for Reddit is more positive than on Twitter, both express similar sentiment changes during key moments of the pandemic and vaccine distribution. This behavior is especially observable as vaccines become widely available to the public, and polarity diminishes. Considering this fact, I feel these results suggest both Twitter and Reddit are valuable data sources that public health officials can utilize for public health surveillance. Interestingly, sentiment appeared to change a bit earlier on Twitter than on Reddit which would be beneficial for quick responses against the backdrop of a disease outbreak.

Twitter is also superior in the ability to access large numbers of Tweets through an API. Though perhaps beneficial for collecting a random sample of users, significant steps need to be taken while cleaning Twitter data to remove bots, news media posts, commercial users, duplicates, and extremely high posting frequencies of some users. It would be interesting for a future study to investigate the sentiment of Twitter news media users and compare it with the results of this study. On the other hand, Reddit appears to be much more valuable for topic modeling and semantic network analysis simply due to the lengths of postings compared to the length of Tweets (i.e., 10k vs 280 characters respectively). The shortened character limit of Tweets most probably contributes to the quick spread of information. However, Reddit users typically take advantage of the much longer character limit and share at times, highly personal stories and experiences related to their health care.

For this reason, Reddit could remain a highly valuable source when considering the development of public health messaging and education campaigns.

Sentiment Driving Factors

Correlating changes in sentiment with developments during the pandemic presents some interesting challenges and ideas alike. The most obvious steep increase in sentiment seems to be correlated with positive news regarding vaccine development, trials, and news of high efficacy, distribution, and availability to those who patiently waited for the vaccine. However, obvious causes of subtle changes remain somewhat elusive without looking deeper into the content. How could the behavior of politicians, celebrities, religious leaders, or an athlete affect sentiment? The dataset was queried for the occurrence of posts mentioning some people in the public eye (PIPE) who had made anti-vaccine statements at public hearings, on social media, or in television interviews.

The Twitter data contained many discussions regarding PIPE but the Reddit data was surprisingly lacking. For example, the Twitter data contained approximately 70k discussions of PIPE who had made negative statements regarding vaccines or who were involved with the antivaccine community and/or in propagating misinformation (e.g., Candace Owens, Joe Rogan, Tucker Carlson, and Phil Valentine). The majority of Twitter sentiment around these discussions was determined to be more negative. Reddit only had 470 mentions of the same people. The sentiment associated with these occurrences was overwhelmingly negative on Twitter and slightly more positive on Reddit. Nonetheless, both communities expressed similar changes in sentiment at key times. Further analysis by the method of topic modeling, or semantic analysis of PIPE-focused discussion as well as news media content would provide a deeper understanding of how the sentiments propagate. These methods could prove to be extremely valuable if implemented on a daily or weekly basis by public health officials. Proper implementation of such techniques would surely provide profound, near real-time insight into the public mind.

For example, a short analysis of Tweets mentioning four news media personalities potentially reveals interesting insights into sentiment-driving factors. For news media personalities (NMP), Joe Rogan (JR) had a total of 6136 Tweets in this dataset. The maximum sentiment for JR occurred in Jan 2021 but quickly reversed to reach a minimum in February 2021. The overall mean sentiment was determined to be 0.295. The mention of JR in a Tweet received the maximum score of 49929 and a mean of 62.11. References to Tucker Carlson (TC) occurred 4843 times in the dataset. Sentiments associated with TC reached a maximum during April 2021 at 0.355, a minimum of 0.074 in August 2020, and a mean of 0.249. Tweets mentioning TC had a maximum of 24586 likes, and a mean of 39.35. PV had 1264 mentioning his name and reached his maximum sentiment of 0.4 in Nov 2021, a minimum of 0.107 in December 2021, and a mean of 0.257. Tweets mentioning Phil Valentine (PV) had a maximum of 12264 likes and a mean of 62.75. Lastly, Candace Owens (CO) reached her maximum sentiment of 0.43 in Jan 2022, a minimum score of 0.038 in October 2021, and a mean of 0.229. Tweets mentioning CO displayed a maximum of 7195 likes, and a mean of 38.60. CO occurred 996 times in the dataset. It is mentionable that occurrences of these personalities in Tweets were not constant throughout the pandemic. TC, and PV both were first mentioned in July 2020, JR in October 2020, and CO in November 2020. Though oscillatory throughout the pandemic, sentiment related to the subgroup of news media personalities was also overall more negative than positive. Moreover, Tweets referencing this group were typically more related to anti-vaccine controversy or death (i.e. death of PV) rather than news about vaccine development. Though the sentiment for the subgroup was overwhelmingly negative on average, some themes become noticeable with inspection of Tweet content. For example, the most liked Tweet associated with JR was "*I love how the same people who don't want us to listen to Joe Rogan, Aaron Rodgers about the covid vaccine, want us to listen to Big Bird & Elmo*", clearly a vaccine-hesitant or anti-vaccine statement.

Notably, it is interesting to compare the number of Tweets with the total number of maximum likes. This combined set of news media personalities had a total of 14017 with 93974 associated highest like count. The high number of likes displayed within these Tweets shows that a much higher number of users are involved in reading Tweets and are therefore potentially influenced by the content. The risk of severe negative health outcomes increases with failure to comply with health-protective behavior recommendations set forth by public health officials, and these findings suggest that polarized messages from societal elites may downplay these risks, unduly contributing to an increase in the spread of COVID-19.

Social Media and Digital Health

Alongside the numerous public health preventive measures (i.e., social distancing, shelter-in-place, stay-at-home orders, lockdowns, quarantine, etc.) implemented to control the spread of the virus, there is general scientific consensus that the COVID-19 vaccine is protective against the SARS-COV-2. However, the spread of misinformation, disinformation, and fake news plays a major role in vaccine hesitancy, low vaccination rates, disease outbreaks as well as morbidities and untimely deaths from vaccine-preventable illnesses. Accordingly, the leverage of textual data obtained from social media platforms could facilitate rapid *and* inexpensive public sentiment analysis thereby enabling the implementation of appropriate messaging, digital interventions, and policies. Digital health technologies and Artificial Intelligence (Shaban-Nejad et al., 2018) are novel, ideal, and effective tools that could facilitate the delivery of accurate, timely, and targeted health information to the general public. For instance, this intervention could be implemented as automated personalized messages and education delivered to individuals based on the content and sentiments from their social media posts. Personalized educational interventions can provide clear, unambiguous recommendations/policies/messages on vaccine safety, efficacy, availability, accessibility, affordability, acceptability, etc. could be impactful. Pivoting online forum discussions on vaccines to accurate and evidence-based information would conceivably facilitate Precision Health Promotion (Shaban-Nejad et al., 2020) and increased health literacy to promote vaccine confidence.

It is widely known that surveillance of population movement and interaction is significant in controlling disease transmission, prompting many countries to focus on digital health technologies capable of recording both movement and relevant environmental biomarkers. These location-based biomarkers range from fine particulate matter in the air to descriptive statistics detailing local access to green space and public transportation (De Brouwer, 2021). Non-genetic biomarkers such as these are reflective of a population’s “exposome”, a public health concept demonstrating the connection between environmental pressures and overall health status (Miller, 2014). Overtime, external pressures from environmental determinants influence an individual's biological index, making disease development and progression unique (Miller, 2014). As misinformation/disinformation leads to vaccine hesitancy in societies around the Earth, public health surveillance methodologies (e.g., topic modeling, semantic network analysis, sentiment analysis) of social media data can be used to improve a population's "digital exposome" (Melton et al., 2021) (Lopez-Campos et al., 2017). Furthermore, these methods can be used in conjunction with informatics from wearable sensor devices, smartphone-based sensors, environmental hyperspectral and remote sensing campaigns, and geolocation technologies. These innovative tools could be employed to investigate exposome complexities and aid in site-specific mitigation plans for smaller community groups (e.g., school board meetings) (Olusanya et al., 2021) and for a larger population where text data can be harvested.

Though a multimodal approach would be beneficial to employ, it is likely that many populations could fall through the gaps due to IT infrastructure. Notably, a major challenge in some African countries is the lack of adequate infrastructure for internet connectivity, power/electricity supply, and EHR management, which are needed for any real-time application for disease surveillance and monitoring. While countries such as Kenya, Libya, and Nigeria have considerably good internet coverage (approximately 80 percent average), others like Madagascar, South Sudan, and Western Sahara have minimal coverage (less than 10 percent) (Statista, 2020). African nations with high internet access are suitably positioned to leverage applications for contact tracing, disease surveillance, data visualization, and vaccine distribution (WHO, 2021) (JHUCRC, 2021). For harder-to-reach/rural areas (without internet access), satellite internet devices and offline digital health strategies can be adopted to collate, integrate and analyze population data (Ogunleye, 2020). It is imperative for countries with densely populated city centers and limited internet connectivity (e.g., Egypt, and South Africa) to improve and stabilize their capacity for implementation. A positive future outlook is the internet coverage which has rapidly increased by approximately 12,000 percent in 2020-2021) within Africa.

Conclusions of results

This section described interpretation of results from topic modeling (LDA), semantic network analysis, and two sentiment analyses from four datasets harvested throughout the COVID-19 pandemic (January 1, 2020 – March 1, 2022). Nine specific questions were stated to test the hypothesis stated in the introduction of this dissertation. To refresh the mind of the reader, the questions and results are stated as followed:

1. How has sentiment changed throughout the pandemic? *Overall, sentiment increased throughout the pandemic until March 2021, and April 2021 for Twitter and Reddit respectively. Sentiment expressed on both platforms then declined relatively throughout the remain months in the Twitter and Reddit 2.0 data set.*
2. Is it possible to correlate changes in sentiment with major events during the pandemic? *Yes, steep increase in sentiment seemed to be correlated with positive news regarding vaccine development, trials, and news of high efficacy,*

distribution, and availability to those who patiently waited for the vaccine. It is highly likely that the gradual decline was related to a combination of unfortunate events related to the pandemic (e.g., misinformation, pandemic fatigue, and falling vaccine efficacy).

3. How does sentiment vary between Reddit communities and Twitter? *In these data, sentiment expressed on Reddit was generally higher than sentiment expressed on Twitter. However, overall sentiment expressed on both platforms appeared to behave closely and shifted relatively similar times. Lastly, changes in sentiment expressed on Reddit tended to follow behind sentiment on Twitter.*

4. Do Reddit and Twitter discuss similar topics at similar times? *Yes, at times certain key terms appeared a similar times (e.g., related to side-effects, boosters, mandates, vaccine development). However, Reddit communities in these data were expressed terms related to side-effect much more than Twitter.*

5. What are commonly discussed topics on Reddit and Twitter related to COVID-19 vaccines? *Related to COVID-19 vaccines specifically, most topics were related to vaccine development, side-effects, boosters, and appointment availability.*

6. Will the occurrence of misinformation be prevalent in data? *Misinformation was detected in both Reddit and Twitter data. However, the occurrence of misinformation related terms were not as prevalent as expected.*

7. Will the occurrence of conspiracy theories be prevalent in data? *Yes, terms related to conspiracy theories were detected but more so in the Twitter dataset (i.e., Bill Gates).*

8. Is one social media platform better for public health surveillance? *Both remain highly useful but the answer to this question depends on the specific use case. The lower character limit, user base, and leading sentiment changes suggest that Twitter could be more efficient in the detection of early disease outbreak. Furthermore, the longer character limit and content detected from the Reddit data suggest that Reddit could be a better source of semantic quantity, and better to use for the development of educational or public health messaging campaigns.*

9. How generalizable is the approach presented in this research? *I strongly feel that this approach is highly generalizable and could be use to investigate further disease outbreaks and other phenomena discussed on social media.*

Major Research Contributions

Contribution to Data Science and NLP

This work contributed to the growing field of data science and NLP in a few specific ways. First of all, to complete this body of work, two large datasets were created from harvesting social media text that total over 13.5 million, as well as a sentiment-labeled dataset consisting of approximately 3600 Tweets and an augmented dataset of approximately 50 thousands Tweets. These data are publicly available with reasonable requests for researchers to further investigate phenomena related to vaccine perception and public discourse during one of the worst pandemics in recent history. Secondly, this work offers code repositories that include techniques for topic modeling, sentiment analysis, back-translation for data augmentation, fine-tuning of several BERT-like models, text harvesting from Reddit and Twitter, graph network building, data querying, and most importantly data cleaning. These repositories are also available at my personal Github account. This work also offers several fine-tuned models (some for testing) that were fine-tuned on a custom-labeled dataset related to COVID-19-related social media posts. These are also available at my huggingface.co repository. Furthermore, similar to many domains of science, a single analysis can offer a researcher some insight into a phenomenon. Many times big assumptions must be made or only an aspect of the phenomenon may be realized. The work presented in this dissertation provides a highly successful example of how a multimodal analysis lends itself to understanding the challenges of fast-moving, fast-changing problems with big data (e.g., a pandemic). Finally, this research has contributed to 11 publications to which I was the lead author on six. These publications have been featured in two high-impact journals and several conferences. One work in particular " Public sentiment analysis and topic modeling regarding COVID-19 vaccines on the Reddit social media platform: A call to strengthen vaccine confidence" has been cited over 52 times.

Public Health Implications

The application of these findings could have momentous impacts on the public health sector in the fight against infectious diseases such as COVID-19. This work provides an example of how data science, and especially NLP can be employed for near real-time public health surveillance and historical pandemic research. Most importantly, the methods and results presented in this work show proof of concept that these developed techniques, utilized in a timely fashion could guide public health officials at local, state, and federal levels to develop not only real-time interventions as well as national education campaigns. The timeframe of analysis is important to consider for future pandemic and disease public health officials. For example, this study provided a high-level view of public discourse from an average monthly perspective. Analysis of larger timescales excels at showing significant changes over time. It is important for future researchers to consider that some analytical subtleties could be potentially lost when dealing with these larger timescales. These subtle changes in sentiment or discourse could prove to be valuable for quick response efforts when attempting to combat misinformation or detect disease outbreaks. On the contrary, for example, daily fluctuations in sentiment can oscillate widely and appear volatile or noisy to a researcher. Conceivably, real-time monitoring of social media discussions could be implemented on smaller time scales such as daily or weekly. However, analysis on daily or weekly timescales should be viewed in concert with monthly or quarterly analysis as well.

In the hands of a front-line public health analyst, the tools used in this research, organized in a simple-to-use Graphical User Interface (GUI) software package could provide early warning to policymakers of outbreaks, misinformation, disinformation, and/or misunderstandings, therefore streamlining the policy implementation process. For a hypothetical example, imagine a public health analyst (using such a GUI) collects data from a major metropolitan area social media community group (Group A) daily and begins to notice some anomaly (i.e., misinformation, anti-vaccine, strange symptoms) in content or sentiment over a few days. In this hypothetical situation, let us imagine the anomaly is a combination of symptoms that are related to some future highly contagious virus first detected in a different metropolitan area that has daily flights to Group A's city. Being a competent analyst in a city with competent/unified leadership in public health and government, these anomalies are reported and proper mitigative policy steps can be put in place quickly to slow and/or stop the spread of the new outbreak. Admittedly, this hypothetical example assumes a best-case scenario in leadership which might not always be the case. Nonetheless, early detection and messaging campaigns could help sway public opinion before other those less informed to have the opportunity to spread harmful misinformation or misunderstood facts.

Challenges and Limitations

This study has some limitations. Additional challenges occur when conducting sentiment analysis in social media text due to long-standing problems with detecting sarcasm, often leading to false positives or false negatives. At least one false positive was detected in a comment thread. One user posted sarcastically, "*They never gave me a bloody sticker*!". Though a human can see the sarcastic intent of such a post, TextBlob rated this post as negative and highly subjective. Moreover, TextBlob usually exhibits modest returns inaccuracy (50–70%), and there may be room for improvement. Reddit is superior to other social media platforms in several ways in user numbers and data quality. Though many outstanding studies have been conducted using Twitter data, it is estimated that approximately 50% of Twitter accounts could be BOTS (Allyn 2020). A recent study by Memon and Carly (2020) reported that up to 14% of COVID-19-related posts on Twitter were composed by BOTS. Though some BOTS exist, the operational design of Reddit community interaction does not lend itself to typical BOT behavior. Nonetheless, the site still is not perfect. Only broad data are available regarding the Reddit user base. While some demographic, financial, gender and geographic data have been gathered (Sattleberg 2019), geotagged posts are not a regular occurrence on most Reddit posts. Moreover, high-resolution demographic data are not available or recorded. This lack of geocoded data makes comparison with specific regional/city-wide polling or surveys impossible unless the subreddit is explicitly based on a geographical community or dedicated to a specific demographic. That being, Reddit data are typically not ideal for studies of a highly specific geographic area or demographic studies.

Other biases could be introduced in data labeling which could confuse a model. For example, many pro-vaccine social media users express extremely negative views and sentiments regarding the anti-vaccine community. How would BERT classify such an occurrence? Though their expressed sentiment is positive towards the vaccine, many NLP algorithms and data labelers would potentially struggle with this type of classification. Even though great care was taken with this study to remove Tweets by BOTs, or Tweets from highly repetitive users from Twitter, and choose unbiased subreddits. It is possible some could have still slipped through the data cleaning process. Moreover, augmented data can potentially cause problems with overfitting when fine-tuning models due to relatively similar semantic content. I limited training epochs and closely monitored the relationship between the training loss and the validation loss to mitigate this potential problem. Future work could involve efforts to create a larger labeled data set that would include not only COVID-19 vaccine sentiment but other vaccines as well.

Future Work

Public Health Usage

Despite the challenges endured over the last 2.5 years during the COVID-19 pandemic, the future is bright for NLP and public health. As more BERT-like models are created, the power of text classification will continue to improve. In the immediate future, further development of low-human-effort surveillance systems optimized for rapid collection of data would allow for real-time analysis of public emotion during times of disease outbreaks. The use of software applications with pretrained models could expand into many subfields of public health as well enhance previous models trained to classify clinical notes (e.g., ClinicalBERT, BioBERT) and other public health data (e.g., VAERS, patient data collection). The logical next steps of this work would be related to understanding the geographical and demographical perception of vaccines and willingness to accept misinformation. Fine-tuning models to extrapolate geographical and demographical differences in sentiment could provide insight into the attitudes of populations at greatest risk of the debilitating outcome. In addition to geo- and demographic-specific data mining, targeting public discourse during times of peak infection, vaccine releases, or celebrity/athlete/political figure deaths due to disease could greatly bolster public health response. The expansion of such disease prediction models based on sentiment detection could also positively influence evidence-informed policy development. Discernment of these dynamic populations could allow public health officials to design personalized policy communication strategies. For example, if a population is hesitant because of some specific misinformation or misunderstanding, specific education or a public marketing campaigns (e.g., "The More You Know", celebrity messaging) could be deployed on television, radio, social media, or digital signs on high-traffic roads or railways. By providing the necessary tools to better understand public emotion related to disease prevention, control, and containment, policymakers would be more well equipped to evaluate program successes and highlight any need for repositioning.

That all being said, significant psychological, sociological, and cultural studies are desperately needed to understand what drives certain populations, news media, politicians, and entertainers to so readily accept and propagate misinformation, and conspiracy theories rather than directly observable facts. Such studies would not only benefit future public health responses but also many other areas of life where misinformation and disinformation have taken hold. The success of digital interventions and education campaigns would likely be limited without a more thorough understanding of how to reach these populations.

Conclusion

The goal of the research presented in this dissertation was to surveil public discourse related to COVID-19 vaccination from large corpora of social media textual data through sentiment analysis, topic modeling, and semantic network analysis. Sentiment analysis through the fine-tuned DistilRoBERTa model revealed that even though Twitter content was more negative on average than content expressed on Reddit, relatively similar changes in sentiment occurred among users of both online platforms and that these major shifts in sentiment occurred with significant developments and events during the pandemic. Temporal LDA topic modeling and semantic network analysis provided insight into how public discourse related to COVID-19 and vaccinations, misinformation, and vaccine hesitancy evolved over the course of 26 months. This work also harvested two large datasets (*Coronavirus Twitter Data and Reddit 2.0 Reddit data*) consisting of over 13.5 million text entries that can be utilized by researchers to further investigate phenomena related to vaccine perception and public discourse during one of the worst pandemics in recent history. Furthermore, this work provides a framework that could be scaled and utilized by public health officials to monitor disease outbreaks in near real-time. Hopefully, the results from this study will help to guide and facilitate the implementation of targeted digital interventions among vaccine-hesitant populations and provide insights to public health officials to inform decision-making and effective policy development.

REFERENCES

Aiello, A. E., Renson, A., & Zivich, P. (2020). Social media-and internet-based disease surveillance for public health. *Annual review of public health*, *41*, 101.

Alessa, A., & Faezipour, M. (2019). Preliminary flu outbreak prediction using twitter posts classification and linear regression with historical centers for disease control and prevention reports: Prediction framework study. *JMIR public health and surveillance*, *5*(2), e12383.

Allyn, B. (2020). Researchers: Nearly half of accounts tweeting about coronavirus are likely bots. *NPR. org [Internet]*, *20*.

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W. H., Jin, D., Naumann, T., & McDermott, M. (2019). Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323*.

Amdahl, G. M. (1967, April). Validity of the single processor approach to achieving large scale computing capabilities. In *Proceedings of the April 18-20, 1967, spring joint computer conference* (pp. 483-485).

Aschwanden, C. (2020). The false promise of herd immunity for COVID-19. *Nature*, 26-28.

Balmas, M. (2014). When fake news becomes real: Combined exposure to multiple news sources and political attitudes of inefficacy, alienation, and cynicism. *Communication research*, *41*(3), 430-454.

Beddiar, D. R., Jahan, M. S., & Oussalah, M. (2021). Data expansion using back translation and paraphrasing for hate speech detection. *Online Social Networks and Media*, *24*, 100153.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, *3*, 993-1022.

Chad A. Melton, Jintae Bae, Olufunto A. Olusanya, Jon Hael Brena, Eun Kyong Shin, and Arash Shaban-Nejad. “Semantic Network and Content Analysis of COVID-19 Vaccine Related Social Media Text”. To appear in Proceedings of the AAAI International Workshop on Health Intelligence (W3PHIAI 2022). Vancouver, Canada. Feb 22- March 1, 2022

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J., & Blei, D. (2009). Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems*, *22*.

Clauset, A., Newman, M. E., & Moore, C. (2004). Finding community structure in very large networks. *Physical review E*, *70*(6), 066111.

De Brouwer, W., Patel, C. J., Manrai, A. K., Rodriguez-Chavez, I. R., & Shah, N. R. (2021). Empowering clinical research in a decentralized world. *NPJ Digital Medicine*, *4*(1), 1-5.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dubey, A. D. (2021). Public Sentiment Analysis of COVID-19 Vaccination Drive in India. *Available at SSRN 3772401*.

Euler, L. (1741). Solutio problematis ad geometriam situs pertinentis. *Commentarii academiae scientiarum Petropolitanae*, 128-140.

Funk, C., & Tyson, A. (2020). Intent to get a COVID-19 vaccine rises to 60% as confidence in research and development process increases. *Pew Research Center*, *3*.

Funk, C., & Tyson, A. (2021). Growing share of Americans say they plan to get a COVID-19 vaccine–or already have. *Pew Research Center*.

Gao, H., Guo, D., Wu, J., Zhao, Q., & Li, L. (2021). Changes of the public attitudes of China to domestic COVID-19 vaccination after the vaccines were approved: a semantic network and sentiment analysis based on sina weibo texts. *Frontiers in Public Health*, *9*, 723015.

Gbashi, S., Adebo, O. A., Doorsamy, W., & Njobeh, P. B. (2021). Systematic Delineation of Media Polarity on COVID-19 Vaccines in Africa: Computational Linguistic Modeling Study. *JMIR medical informatics*, *9*(3), e22916.

Ghosh, D., & Guha, R. (2013). What are we 'tweeting' about obesity? Mapping tweets with topic modeling and Geographic Information System. *Cartography and geographic information science*, *40*(2), 90-102.

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, *1*(12), 2009.Bermingham, A., & Smeaton, A. F. (2010, October). Classifying sentiment in microblogs: is brevity an advantage?. In *Proceedings of the 19th ACM international conference on Information and knowledge management* (pp. 1833-1836).

Google (2020). E-Conomy Africa 2020. Africa's $180 billion internet economy future. Retrieved from https://www.ifc.org/wps/wcm/connect/e358c23f-afe3-49c5-a509-034257688580/e-Conomy-Africa-2020.pdf?MOD=AJPERES&CVID=nmuGYF

Jang, H., Rempel, E., Roth, D., Carenini, G., & Janjua, N. Z. (2021). Tracking COVID-19 discourse on twitter in North America: Infodemiology study using topic modeling and aspect-based sentiment analysis. *Journal of medical Internet research*, *23*(2), e25431.

Jeon, D., Lee, J., & Kim, C. (2022). User Guide for KOTE: Korean Online Comments Emotions Dataset. *arXiv preprint arXiv:2205.05300*.

Johns Hopskins University Coronavirus Resource Center. (2021). COVID-19 Dashboard. https://coronavirus.jhu.edu/map.html

Krishnan, G. S., Sowmya Kamath, S., & Sugumaran, V. (2021, June). Predicting vaccine hesitancy and vaccine sentiment using topic modeling and evolutionary optimization. In *International Conference on Applications of Natural Language to Information Systems* (pp. 255-263). Springer, Cham.

Ledford, H. (2021). COVID vaccines and blood clots: five key questions. *Nature*, *592*(7855), 495-496.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics, 36*(4), 1234-1240.

Lopez-Campos, G., Merolli, M., & Martin-Sanchez, F. (2017). Biomedical informatics and the digital component of the exposome. In *MEDINFO 2017: Precision Healthcare through Informatics* (pp. 496-500). IOS Press.

Loria, S. (2018). textblob Documentation. *Release 0.15*, *2*(8).

Luo, C., Chen, A., Cui, B., & Liao, W. (2021). Exploring public perceptions of the COVID-19 vaccine online from a cultural perspective: Semantic network analysis of two social media platforms in the United States and China. *Telematics and Informatics*, *65*, 101712.

Mahase, E. (2020). Covid-19: Moderna vaccine is nearly 95% effective, trial involving high risk and elderly people shows. *BMJ: British Medical Journal (Online)*, *371*.

Mantas, J. (2020). Application of topic modeling to tweets as the foundation for health disparity research for COVID-19. *The Importance of Health Informatics in Public Health during a Pandemic*, *272*, 24.

Mattei, M., Caldarelli, G., Squartini, T., & Saracco, F. (2021). Italian Twitter semantic network during the Covid-19 epidemic. *EPJ Data Science*, *10*(1), 47.

Melton, C. A., Olusanya, O. A., Ammar, N., & Shaban-Nejad, A. (2021). Public sentiment analysis and topic modeling regarding COVID-19 vaccines on the Reddit social media platform: A call to action for strengthening vaccine confidence. *Journal of Infection and Public Health*, *14*(10), 1505-1512.

Melton, C., Olusanya, O. A., & Shaban-Nejad, A. (2021). Network analysis of COVID-19 vaccine misinformation on social media. *Stud Health Technol Inform*, *287*, 165-166.

Melton, C.A, White, BM., Davis, RL., Bednarczyk, RA., Shaban-Nejad, A. “Sentiment Analysis of COVID-19 Vaccines: A Comparative Study of Twitter and Reddit Social Media Platforms”. Journal of Medical Internet Research 2022 (*in review*).

Memon, S. A., & Carley, K. M. (2020). Characterizing covid-19 misinformation communities using a novel twitter dataset. *arXiv preprint arXiv:2008.00791*.

Miller, G. W., & Jones, D. P. (2014). The nature of nurture: refining the definition of the exposome. *Toxicological sciences*, *137*(1), 1-2.

Ogunleye, O. O., Basu, D., Mueller, D., Sneddon, J., Seaton, R. A., Yinka-Ogunleye, A. F.& Godman, B. (2020). Response to the novel corona virus (COVID-19) pandemic across Africa: successes, challenges, and implications for the future. *Frontiers in pharmacology*, 1205.

Olusanya OA, White B, Amuchi B, Melton CA, and Shaban-Nejad A. Perceptions and Misinformation on COVID-19 Mask Mandate During Tennessee School Board Meetings: Analyses and Recommendations for In-Person Learning. Submitted to Journal of *Qualitative Health Research*. **(Under Review)**

Olusanya OA, White B, Melton CA, Shaban-Nejad A. Examining the Implementation of Digital Health to Strengthen the COVID-19 Pandemic Response and Recovery and Scale up Equitable Vaccine Access in African Countries. JMIR Form Res. 2022 May 17;6(5):e34363. doi: 10.2196/34363. PMID: 35512271; PMCID: PMC9116456.

Park, S., & Park, J. (2021). Identifying the Knowledge Structure and Trends of Outreach in Public Health Care: A Text Network Analysis and Topic Modeling. *International journal of environmental research and public health*, *18*(17), 9309.

Praveen, S. V., Ittamalla, R., & Deepak, G. (2021). Analyzing the attitude of Indian citizens towards COVID-19 vaccine–A text analytics study. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, *15*(2), 595-599.

Puri, N., Coomes, E. A., Haghbayan, H., & Gunaratne, K. (2020). Social media and vaccine hesitancy: new updates for the era of COVID-19 and globalized infectious diseases. *Human vaccines & immunotherapeutics*, *16*(11), 2586-2593.

Ritonga, M., Al Ihsan, M. A., Anjar, A., & Rambe, F. H. (2021, February). Sentiment analysis of COVID-19 vaccine in Indonesia using Naïve Bayes Algorithm. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1088, No. 1, p. 012045). IOP Publishing.

Rutjens, B. T., van der Linden, S., & van der Lee, R. (2021). Science skepticism in times of COVID-19. *Group Processes & Intergroup Relations*, *24*(2), 276-283.

Sanders, A. C., White, R. C., Severson, L. S., Ma, R., McQueen, R., Paulo, H. C. A., ... & Bennett, K. P. (2021). Unmasking the conversation on masks: Natural language processing for topical sentiment analysis of COVID-19 Twitter discourse. *medRxiv*, 2020-08.

Santoveña-Casal, S., Gil-Quintana, J., & Ramos, L. (2021). Digital Citizens' Feelings in National# Covid19 Campaigns in Spain.

Sattelberg, W. (2019). The demographics of Reddit: Who uses the site. *Tech Junkie*.

Shaban-Nejad, A., Michalowski, M., & Buckeridge, D. L. (2018). Health intelligence: how artificial intelligence transforms population and personalized health. *NPJ digital medicine*, *1*(1), 1-2.

Shaban-Nejad, A., Michalowski, M., Peek, N., Brownstein, J.S. and Buckeridge, D.L., 2020. Seven pillars of precision digital health and medicine.

Shin, E. K., Choi, H. Y., & Hayes, N. (2021). The anatomy of COVID-19 comorbidity networks among hospitalized Korean patients. *Epidemiology and Health*, *43*.

Srinivasa-Desikan, B. (2018). *Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras*. Packt Publishing Ltd.

Statista (2020, December). Share of internet users in Africa as of December 2020, by country. Retrieved from https://www.statista.com/statistics/1124283/internet-penetration-in-africa-by-country.

US Food and Drug Administration, & approves first COVID, F. D. A. (19). vaccine. 2021.

van der Linden, S., Dixon, G., Clarke, C., & Cook, J. (2021). Inoculating against COVID-19 vaccine misinformation. *EClinicalMedicine*, *33*.

Wang, H., Can, D., Kazemzadeh, A., Bar, F., & Narayanan, S. (2012, July). A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In *Proceedings of the ACL 2012 system demonstrations* (pp. 115-120).

Wojcik, S., & Hughes, A. (2019). Sizing up Twitter users. *PEW research center*, *24*.

World Health Organization. (2014). Report of the Sage working group on vaccine hesitancy.

World Health Organization. (2021). *COVID-19 and digital health: What can digital health offer for COVID-19?* World Health Organization. https://www.who.int/china/news/feature-stories/detail/covid-19-and-digital-health-what-can-digital-health-offer-for-covid-19.

World Health Organization. (2021). WHO covid-19 dashboard. https://covid19.who.int/cdc

Wu, W., Lyu, H., & Luo, J. (2021). Characterizing Discourse about COVID-19 Vaccines: A Reddit Version of the Pandemic Story. *arXiv preprint arXiv:2101.06321*.

Xenouleas, S., Tsoukara, A., Panagiotakis, G., Chalkidis, I., & Androutsopoulos, I. (2022). Realistic Zero-Shot Cross-Lingual Transfer in Legal Topic Classification. *arXiv preprint arXiv:2206.03785*.

Yin, F., Wu, Z., Xia, X., Ji, M., Wang, Y., & Hu, Z. (2021). Unfolding the determinants of COVID-19 vaccine acceptance in China. *Journal of medical Internet research*, *23*(1), e26089.

Yoo, S. Y., & Lim, G. G. (2021). Analysis of news agenda using text mining and semantic network analysis: Focused on COVID-19 emotions. *Journal of Intelligence and Information Systems*, *27*(1), 47-64.

Zhang, J., Featherstone, J. D., Calabrese, C., & Wojcieszak, M. (2021). Effects of fact-checking social media vaccine misinformation on attitudes toward vaccines. *Preventive Medicine*, *145*, 106408.

APPENDIX

Table 1: Dataset characteristics

Dataset Name	Dates	Harvest count	Final count	Unique authors
Reddit 1.0	December 1, 2020 to May 16, 2021	18000	11641	8021
Reddit 1.2	January 1, 2020 to October 31, 2021	300000	31432	20429
Reddit 2.0	January 1, 2020 to March 1, 2022	579241	69079	9933
Twitter	January 1, 2020 to March 1, 2022	13000000	9518270	3006075

Table 2: DistilRoBERTa Fine-tuning Training Metrics

Training Loss	Epoch	Step	Validation Loss	Precision	Accuracy	F1
0.5903	0.4	500	0.4695	0.7342	0.7728	0.7890
0.3986	0.8	1000	0.3469	0.8144	0.8596	0.8684
0.2366	1.2	1500	0.1939	0.9313	0.9260	0.9253
0.1476	1.6	2000	0.1560	0.9207	0.9452	0.9465
0.1284	2.0	2500	0.1167	0.9561	0.9592	0.9592

Table 3: Reddit 1.0 Latent Topics

Topic #	Latent Topics
	Combined Data Set (December 1, 2020 - December 31, 2020)
1	vaccine , people , effect , time , many , thing , year , death , month , good
2	vaccine , effect , side , week , hour , day , second , fever , symptom , sore
3	vaccine , people , dose , mask , thing , group , datum , year , immunity , efficacy
4	vaccine , virus , people , immune , system , year , antibody , vaccination , immunity , body
5	vaccine , question , contact , concern , action , people , source , news , moderator , answer
	December 2020
1	vaccine, virus, immune, system, question, cell, protein, infection, symptom, body
2	vaccine, dose, trial, group, first, efficacy, datum, case, day, participant
3	vaccine, people, year, thing, effect, time, virus, long, side, good
	January 2021
1	vaccine, dose, effect, people, side, day, second, week, first, shot
2	vaccine, people, virus, year, time, good, immunity, immune, risk, case
	February 2021
1	vaccine, dose, second, effect, day, side, week, people, first, hour
2	vaccine, people, virus, immune, vaccination, immunity, time, antibody, cell, mask
	March 2021
1	vaccine, people, virus, mask, year, thing, immunity, good, time, immune
2	vaccine, vaccination, dose, death, question, effect, week, day, people, concern
	April 2021
1	vaccine, people, mask, thing, year, effect, time, vaccination, virus, death
2	vaccine, people, virus, vaccination, immune, t, immunity, death, effect, time
	May 2021
1	vaccine, people, effect, side, time, second, shot, week, death, day
2	vaccine, people, mask, virus, vaccination, immunity, risk, year, thing, t

Table 4: Reddit 1.0 Latent Topics

Topic #	Word
	Negative
1	vaccine, business, people, fucking, fuck, government, mask, health, treatment, free,
2	vaccine, second, day, side, effect, symptom, dose, shot, week, hour,
3	vaccine, vaccination, shot, today, itâ€, month, response, day, state, time
4	vaccine, part, concern, contact, question, appointment, action, resource, employee, helpful,
5	vaccine, immune, system, body, cold, efficacy, different, variant, cell, term,
6	vaccine, long, virus, effect, term, people, immune, infection, risk, science,
7	people, vaccine, vaccination, virus, thing, t, mask, immunity, stupid, sick
	Neutral
1	vaccine, people, test, antibody, other, part, mask, case, re, body
2	vaccine, virus, nerve, thing, today, week, physician, doctor, couple, different
3	vaccine, vaccination, shot, today, itâ€, month, response, day, state, time
4	people, vaccine, reaction, shot, reason, fever, today, pfizer, vaccination, itâ€
5	vaccine, t, work, life, sore, time, different, story, situation, period
6	vaccine, effect, side, people, second, pfizer, shot, study, year, tomorrow
7	vaccine, immunity, rate, efficacy, herd, virus, immune, video, issue, link
8	vaccine, immune, day, cell, system, people, thing, research, site, month
	Positive
1	vaccine, effect, side, dose, second, day, hour, reaction, death, week
2	people, vaccine, year, good, thing, time, article, mask, shit, population
3	immune, virus, system, cell, antibody, body, vaccine, protein, response, immunity
4	vaccine, doctor, time, trial, right, country, good, link, year, first
5	vaccine, good, needle, doctor, effective, thing, little, moderna, flood, t
6	vaccine, people, many, immunity, year, safe, virus, shot, death, effect
7	vaccine, people, long, time, term, year, virus, thing, many, effect
8	vaccine, test, risk, trial, woman, infection, pregnant, study, family, people

Table 5: Reddit 2.0 Top betweenness central nodes

20-Feb		20-Mar		20-Apr		20-May		20-Jun	
Node	Centrality	Node	Centrality	Node	Centrality	Node	Centrality	Node	Centrality
people	0.886243	people	0.977273	people	0.75764843	people	0.690122	county	0.658339
cases	0.373016	masks	0.060606	need	0.11864629	cases	0.163288	covid19	0.362731
flu	0.206349	health	0.060606	new	0.05711421	deaths	0.107036	paties	0.201991
virus	0.206349	home	0.060606	virus	0.04135024	going	0.08438	cases	0.127786
days	0.198413	m	0.060606	information	0.03570646	f	0.053489	information	0.052632
20-Jul		20-Aug		20-Sep		20-Oct		20-Nov	
Node	Centrality	Node	Centrality	Node	Centrality	Node	Centrality	Node	Centrality
mr	0.468795	cases	0.459695	covid19	0.334713	people	0.517012	vaccine	0.463758
county	0.374193	county	0.252723	people	0.265462	paties	0.202589	people	0.333681
people	0.33707	covid19	0.215686	couy	0.166878	vaccine	0.183526	m	0.119952
cases	0.277505	male	0.17756	health	0.12001	covid19	0.130604	cases	0.11553
covid19	0.195287	female	0.149237	vaccine	0.10804	like	0.094442	virus	0.081981
20-Dec		21-Jan		21-Feb		21-Mar		21-Apr	
Node	Centrality	Node	Centrality	Node	Centrality	Node	Centrality	Node	Centrality
vaccine	0.790704225	vaccine	0.392361	vaccine	0.496097	vaccine	0.482418	vaccine	0.362386
m	0.166827632	doses	0.200452	county	0.159794	shot	0.10924	shot	0.175872
people	0.155305164	county	0.180369	total	0.08593	people	0.101514	people	0.08463
cases	0.113829645	total	0.153824	m	0.072666	new	0.095858	m	0.06979
days	0.106036217	new	0.110287	immune	0.063585	m	0.076774	days	0.061929
21-May		21-Jun		21-Jul		21-Aug		21-Sep	
Node	Centrality	Node	Centrality	Node	Centrality	Node	Centrality	Node	Centrality
vaccine	0.371508	vaccine	0.334411	vaccine	0.512088	vaccine	0.514286	vaccine	0.485507
people	0.164159	people	0.118134	people	0.107698	people	0.096657	people	0.117955
shot	0.110992	shot	0.10318	vaccinated	0.061056	vaccinated	0.088388	including	0.078276
day	0.08498	couy	0.089772	day	0.056043	shot	0.056204	shot	0.075622
vaccinated	0.055578	m	0.06159	shot	0.04509	day	0.042034	covid	0.062854
21-Oct		21-Nov		21-Dec		22-Jan		22-Feb	
Node	Centrality	Node	Centrality	Node	Centrality	Node	Centrality	Node	Centrality
vaccine	0.241747	vaccine	0.364944	vaccine	0.296053	cases	0.329593	people	0.413338
vaccinated	0.1557	people	0.217826	people	0.280222	covid	0.265788	peak	0.28101
fully	0.09848	vaccinated	0.166677	covid	0.176362	people	0.153051	cases	0.226465
shot	0.07618	shot	0.085868	booster	0.125368	day	0.093247	vaccine	0.14582
got	0.056463	like	0.047621	cases	0.121072	vaccine	0.081507	covid	0.130853

Table 6: Twitter Betweenness Centrality

20-Feb		20-Mar		20-Apr		20-May		20-Jun	
Node	Centrality	Node	Centrality	Node	Centrality	Node	Centrality	Node	Centrality
covid19	0.66666667	vaccine	0.51674648	vaccine	0.68124884	vaccine	0.78338583	vaccine	0.93333333
persists	0.5	covid19	0.34145238	covid19	0.23455916	covid19s	0.28086253	covid	0.33333333
vaccine	0.5	vaccines	0.17882803	coronavirus	0.1821511	covid	0.1273824	people	0
flu	0	cough	0.16510827	migvax	0.17629966	world	0.11066625	amp	0
influenza	0	amp	0.12924966	virus	0.0889196	like	0.09074573	covid19	0
20-Jul		20-Aug		20-Sep		20-Oct		20-Nov	
Node	Centrality	Node	Centrality	Node	Centrality	Node	Centrality	Node	Centrality
vaccine	0.71818182	vaccine	0.723809524	vaccine	0.78204829	vaccine	0.511111111	vaccine	0.808322891
covid	0.16363636	covid	0.114285714	covid	0.12629855	covid	0.322222222	covid	0.159200084
covid19	0.05454545	covid19	0.080952381	long	0.09569402	covid19	0.022222222	covid19	0.016760652
vaccines	0.00909091	vaccines	0.004761905	study	0.08758865	vaccines	0.011111111	vaccines	0.00161863
moderna	0	willing	0	covid19	0.08379602	amp	0	canada	0.00093985
20-Dec		21-Jan		21-Feb		21-Mar		21-Apr	
Node	Centrality	Node	Centrality	Node	Centrality	Node	Centrality	Node	Centrality
vaccine	0.72109637	vaccine	0.691797797	vaccine	0.727337093	vaccine	0.65963595	vaccine	0.705557543
covid	0.236943062	covid	0.279742547	covid	0.24619152	covid	0.3021266	covid	0.270435301
tweetstorm	0.029519071	vaccines	0.026444619	johnson	0.029493525	covid19	0.04027713	covid19	0.048580799
covid19	0.022063755	new	0.016129032	covid19	0.011163847	austiexas	0.02402337	long	0.015384615
vaccines	0.000451446	eu	0.016129032	fda	0.002307853	reallyopen	0.02402337	vaccines	0.004951302
21-May		21-Jun		21-Jul		21-Aug		21-Sep	
Node	Centrality	Node	Centrality	Node	Centrality	Node	Centrality	Node	Centrality
vaccine	0.564197531	vaccine	0.514335332	covid	0.562348789	covid	0.560897436	covid	0.558730499
covid	0.40893592	covid	0.45623938	vaccine	0.419961056	vaccine	0.434418146	vaccine	0.425035781
vaccines	0.027219283	covid19	0.064772614	vaccines	0.011241329	vaccines	0.010167183	vaccines	0.022833476
covid19	0.012637068	vaccines	0.035849575	people	0.003921139	covid19	0.00355147	pregna	0.01459854
long	0.012345679	booster	0.017241379	vaccinated	0.001115979	people	0.001825632	covid19	0.005343853
21-Oct		21-Nov		21-Dec		22-Jan		22-Feb	
Node	Centrality	Node	Centrality	Node	Centrality	Node	Centrality	Node	Centrality
covid	0.58918948	covid	0.55753479	covid	0.74877594	covid	0.64300936	covid	0.722121395
vaccine	0.3879491	vaccine	0.416248497	vaccine	0.213942	vaccine	0.34260163	vaccine	0.276567856
vaccines	0.0202975	vaccines	0.02143341	vaccines	0.02511483	vaccines	0.0279256	deaths	0.034482759
big	0.01190476	lamb	0.016122477	people	0.00479403	available	0.0240946	immunity	0.034482759
people	0.0022859	marcus	0.016122477	vaccinated	0.0005988	covid19	0.01718157	vaccines	0.007360355

Table 7: Subreddits for each dataset

Reddit 1.0	Reddit 1.2	Reddit 2.0	Reddit 2.0 Cont	Reddit 2.0 Cont
Vaccines	CoronaVirus	China_Flu	covidlonghaulers	CoronavirusGA
CovidVaccine	COVID19	CoronavirusCanada	AskDocs	CoronavirusSC
CovidVaccinated	VACCINES	CoronavirusNewYork	COVID19_support	CoronavirusNJ
AntiVaxxers	conspiracy_commons	Coronavirus_PH	CoronavirusCA	coronavirusnewmexico
vaxxhappened	COVID19_support	Coronavirus_KY	CoronavirusWA	CoronavirusNE
antivaccine	AntiVaxxers	Coronavirus_NC	CoronavirusOH	CoronavirusIllinois
conspiracy	CovidVaccine	Coronavirus_BC	CoronavirusCanada	CoronavirusIdaho
conspiracytheories	conspiracy	Coronavirus_Ireland	CoronavirusOC	Coronavirushawaii
NoNewNormal	conspiracytheories	COVID19_Ohio	CoronavirusAZ	CoronavirusArmy
conspiracy_commons	COVID19positive	COVID19_Maine	CoronaVirus	CoronavirusLouisiana
COVID19	China_Flu	COVID19_Testimonials	CoronaVirusTX	CoronavirusNV
COVID	CovidVaccinated	COVID19positive	CoronaVirusUT	Coronavirusnevada
coronavirus		modernavaccine	Coronavirus_NZ	CoronavirusMO
		COVID19	CoronavirusTN	CoronavirusMontreal
		Coronavirus	CoronavirusAlabama	CoronavirusMontana
		CoronavirusUS	CoronavirusAL	CoronavirusMN
		HealthAnxiety	CoronavirusMichigan	CoronavirusWI
		CoronavirusUK	CoronavirusMissouri	CoronavirusPA
		COVID19_support	CoronavirusMA	CoronavirusNH
		PfizerVaccine	CoronavirusVA	CoronavirusKansas
		Covid19VaccineRats	CoronaVirusWV	
		GotTheVaccine	CoronavirusOR	
		CovidAnxiety	CoronavirusOregon	
		COVIDVaccineTalk	CoronavirusAZ	

Table 8: LDA Reddit 2.0 negative classified topics

Topic	Words
	20-Jan
0	vaccine, t, people, flu, effect, risk, drug, year, symptom, don
1	vaccine, people, t, study, vaccinated, year, doctor, thing, good, need
2	people, system, immune, virus, vaccine, wuhan, time, good, anxiy, healthy
3	vaccine, people, thing, t, year, problem, need, day, study, vaccinated
4	people, vaccine, t, virus, good, time, flu, shot, disease, need
5	people, lot, year, sick, virus, doctor, flu, system, t, disease
6	people, anti, hospital, coronavirus, wuhan, system, good, t, drug, china
7	doctor, vaccine, year, anti, vaccination, effect, t, thing, child, unvaccinated
	20-Feb
0	people, case, vaccine, time, china, patient, country, sar, immune, virus
1	people, vaccine, flu, time, t, virus, year, day, thing, need
2	virus, vaccine, people, good, hospital, work, mer, time, cell, life
3	people, immune, vaccine, t, system, work, year, flu, don, virus
4	case, people, t, day, virus, time, death, system, shot, flu
5	people, virus, vaccine, year, time, flu, case, symptom, good, pain
6	virus, vaccine, case, flu, day, people, thing, infection, season, time
7	vaccine, flu, virus, good, day, bill, people, vitamin, person, t
	20-Mar
0	people, virus, vaccine, case, time, patient, mask, thing, day, system
1	people, day, mask, virus, time, case, week, work, immunity, good
2	vaccine, test, study, people, patient, case, t, day, virus, social
3	people, vaccine, virus, need, immune, t, system, good, time, thing
4	people, flu, vaccine, day, good, virus, case, time, t, symptom
5	people, t, test, vaccine, time, immune, number, day, cough, coronavirus
6	t, people, day, time, immune, sick, virus, home, system, thing
7	risk, day, work, people, immune, t, week, virus, year, flu
	20-Apr
0	people, test, x, need, t, vaccine, good, week, antibody, thing
1	people, vaccine, virus, t, test, need, day, thing, good, case

Table 8 Continued

Topic	Words
2	people, t, x, need, time, day, test, state, case, testing
3	people, t, need, testing, x, new, virus, test, state, death
4	vaccine, people, t, antibody, test, testing, case, time, x, need
5	people, day, virus, x, test, need, time, case, state, thing
6	vaccine, people, virus, testing, week, case, immunity, day, test, work
7	people, virus, t, work, home, death, x, new, time, test
20-May	
0	test, people, antibody, t, case, time, positive, day, testing, week
1	people, t, time, good, testing, vaccine, thing, need, x, day
2	people, vaccine, t, virus, work, time, day, x, week, case
3	people, t, vaccine, need, immunity, day, work, case, x, virus
4	test, day, people, x, testing, time, vaccine, state, work, need
5	t, people, time, vaccine, mask, day, x, lot, low, right
6	people, t, test, antibody, testing, virus, day, time, x, vaccine
7	vaccine, people, t, need, case, week, virus, day, work, x
20-Jun	
0	case, vaccine, feme, county, death, t, people, x, good, day
1	county, feme, vaccine, death, case, t, virus, people, x,
2	virus, mask, case, week, vaccine, test, antibody, x, patient, people
3	people, vaccine, day, test, mask, time, virus, t, antibody, work
4	people, vaccine, t, thing, good, need, state, time, test, x
5	t, case, people, x, antibody, test, cell, patient, long, number
6	case, t, people, day, covid, symptom, week, virus, time, test
7	people, vaccine, t, good, time, x, mask, virus, covid, effect
20-Jul	
0	people, vaccine, t, mask, case, thing, time, virus, day, covid
1	vaccine, day, people, virus, mask, immune, time, case, t, infection
2	people, day, case, immune, test, virus, t, positive, time, vaccine
3	mask, t, people, case, vaccine, x, thing, virus, new, time

Table 8 Continued

Topic	Words
4	case, people, t, test, death, day, county, week, mask, immunity
5	people, t, vaccine, don, school, time, year, kid, day, virus
6	people, case, infection, t, day, time, immunity, virus, low, patient
7	people, mask, day, vaccine, virus, t, immune, case, covid, test
	20-Aug
0	test, vaccine, case, day, people, immune, system, mask, child, covid
1	t, good, time, people, day, virus, week, immune, thing, vaccine
2	vaccine, day, test, thing, time, people, good, virus, t, life
3	people, t, vaccine, virus, test, antibody, case, immune, time, testing
4	case, county, people, mask, feme, new, t, death, day, good
5	day, virus, thing, x, test, vaccine, t, people, case, week
6	vaccine, t, case, people, sar, new, cell, x, covid, antibody
7	people, vaccine, t, mask, time, case, thing, year, virus, month
	20-Sep
0	t, vaccine, people, day, symptom, covid, month, risk, vitamin, immune
1	|, vaccine, people, virus, test, time, day, week, case, immune
2	|, vaccine, people, t, week, test, covid, time, vitamin, case
3	vaccine, t, good, doctor, mask, year, people, covid, disease, vitamin
4	people, county, case, antibody, mask, test, immunity, patient, day, feme
5	vaccine, people, t, day, case, thing, long, covid, time, new
6	vaccine, people, virus, day, case, risk, vitamin, |, good, patient
7	covid, t, day, time, people, vaccine, mask, thing, flu, shot
	20-Oct
0	vaccine, people, mask, covid, t, thing, virus, year, flu, time
1	vaccine, t, people, year, patient, flu, time, need, immune, bad
2	vaccine, t, people, immune, time, virus, day, case, mask, infection
3	vaccine, people, year, good, t, way, time, effect, shot, thing
4	vaccine, test, case, time, antibody, thing, day, virus, people, datum
5	shot, vaccine, day, immune, school, t, arm, virus, people, flu

Table 8 Continued

Topic	Words
6	vaccine, people, time, t, year, covid, immune, virus, good, system
7	vaccine, t, people, time, day, month, week, thing, year, case
	20-Nov
0	vaccine, people, t, case, covid, day, virus, efficacy, time, good
1	vaccine, people, virus, day, time, covid, case, need, test, week
2	vaccine, case, people, long, covid, virus, good, time, help, need
3	vaccine, t, covid, pfizer, people, year, new, moderna, effect, long
4	vaccine, people, good, year, news, pfizer, need, t, effective, day
5	vaccine, people, covid, long, time, year, case, virus, mask, way
6	vaccine, case, people, t, covid, day, virus, test, time, year
7	vaccine, people, t, day, covid, work, time, week, virus, immune
	20-Dec
0	vaccine, people, t, covid, safe, year, week, pfizer, thing, vaccination
1	vaccine, people, t, covid, day, time, long, need, year, don
2	vaccine, people, covid, t, time, week, day, case, work, long
3	vaccine, week, people, t, shot, month, time, work, day, case
4	vaccine, people, day, thing, year, new, time, good, need, care
5	vaccine, people, covid, long, t, day, good, virus, hospit, care
6	vaccine, people, time, day, week, covid, hospit, test, pfizer, month
7	vaccine, covid, day, people, case, t, today, pfizer, effect, time
	21-Jan
0	vaccine, people, day, county, covid, effect, feme, case, arm, bad
1	vaccine, people, day, dose, county, shot, covid, t, state, vaccination
2	vaccine, day, people, t, shot, time, arm, covid, week, good
3	vaccine, t, people, day, week, new, good, dose, time, work
4	vaccine, day, vaccination, week, dose, people, moderna, thing, time, work
5	vaccine, |, day, shot, week, covid, county, thing, people, t
6	day, vaccine, week, dose, symptom, people, time, covid, effect, arm
7	day, vaccine, covid, shot, time, people, t, work, symptom, year

Table 8 Continued

Topic	Words
	21-Feb
0	vaccine, day, shot, t, arm, dose, week, work, pain, post
1	day, dose, vaccine, people, past, covid, tot, population, vaccinated, t
2	vaccine, shot, week, day, t, dose, arm, people, case, year
3	shot, vaccine, county, vaccination, people, case, covid, day, t, arm
4	vaccine, day, |, week, people, time, t, good, covid, county
5	vaccine, people, t, week, shot, time, covid, effect, good, new
6	vaccine, people, covid, t, shot, time, dose, day, vaccination, hour
7	vaccine, shot, day, county, covid, people, effect, time, symptom, moderna
	21-Mar
0	vaccine, t, people, day, time, covid, thing, effect, good, shot
1	vaccine, people, day, covid, time, t, vaccinated, appointment, shot, week
2	day, shot, dose, week, vaccine, pfizer, effect, time, people, moderna
3	shot, vaccine, day, people, vaccinated, dose, year, past, old, t
4	vaccine, shot, people, arm, day, t, effect, covid, hour, good
5	vaccine, day, people, week, shot, t, time, bad, vaccinated, year
6	vaccine, day, shot, covid, week, t, pfizer, bad, moderna, people
7	shot, vaccine, day, people, t, post, vaccination, today, effect, arm
	21-Apr
0	shot, vaccine, t, effect, day, pain, time, covid, people, week
1	vaccine, symptom, t, covid, shot, day, case, week, time, bad
2	vaccine, day, bad, t, covid, anxiy, shot, year, good, norm
3	shot, day, people, vaccine, |, arm, week, t, bad, covid
4	vaccine, day, shot, people, effect, vaccinated, dose, pfizer, week, arm
5	vaccine, week, day, shot, effect, people, covid, t, today, pfizer
6	vaccine, shot, day, vaccinated, covid, people, t, time, moderna, week
7	vaccine, day, shot, people, good, arm, bad, fever, hour, appointment
	21-May
0	vaccine, t, people, shot, covid, week, day, vaccinated, pfizer, symptom
1	vaccine, shot, day, people, vaccinated, dose, time, pfizer, population, effect

Table 8 Continued

Topic	Words
2	vaccine, people, day, week, work, month, t, shot, good, Pfizer
3	vaccine, shot, day, covid, week, t, effect, time, vaccinated, people
4	vaccine, day, shot, covid, vaccinated, people, bad, effect, week, good
5	|, shot, vaccine, day, covid, people, vaccinated, t, lot, year
6	vaccine, people, day, effect, vaccinated, covid, week, shot, bad, period
7	vaccine, shot, day, symptom, t, people, bad, headache, effect, week
	21-Jun
0	vaccine, people, shot, t, day, covid, time, vaccinated, vaccination, week
1	vaccine, shot, day, people, time, week, vaccinated, covid, effect, post
2	vaccine, day, shot, vaccinated, effect, pfizer, week, people, year, covid
3	vaccine, day, pfizer, week, heart, moderna, shot, symptom, doctor, work
4	vaccine, day, covid, shot, vaccinated, people, t, week, effect, arm
5	vaccine, day, shot, t, people, effect, week, pain, time, symptom
6	day, people, vaccinated, shot, vaccine, t, population, dose, year, covid
7	day, vaccine, week, people, vaccinated, month, pfizer, period, effect, shot
	21-Jul
0	vaccine, covid, day, shot, people, t, symptom, good, time, bad
1	vaccine, people, covid, case, vaccinated, day, delta, t, effect, variant
2	vaccine, people, covid, pfizer, vaccinated, week, effect, symptom, shot, day
3	vaccine, day, vaccinated, people, shot, covid, symptom, bad, pain, t
4	vaccine, people, covid, t, effect, vaccinated, pfizer, don, doctor, day
5	day, week, vaccine, people, shot, t, covid, long, vaccinated, effect
6	vaccine, day, shot, vaccinated, people, mask, t, covid, month, effect
7	vaccine, day, vaccinated, people, shot, dose, pain, population, year, effect
	21-Aug
0	vaccine, shot, day, covid, vaccinated, week, t, year, pfizer, people
1	vaccine, shot, day, covid, effect, t, people, vaccinated, pfizer, moderna
2	vaccine, people, t, covid, day, vaccinated, shot, bad, time, month
3	day, vaccine, post, coronavirusca, r, rule, covid, effect, vaccinated, contact

Table 8 Continued

Topic	Words
4	vaccine, vaccinated, people, covid, effect, time, day, shot, case, week
5	vaccine, vaccinated, day, people, shot, vaccination, mask, virus, dose, |
6	day, vaccinated, vaccine, covid, people, shot, case, population, time, hospit
7	|, vaccine, vaccination, people, vaccinated, day, doctor, pfizer, shot, year
	21-Sep
0	shot, day, vaccine, vaccinated, effect, t, people, covid, good, bad
1	vaccine, covid, shot, people, day, effect, week, booster, time, t
2	vaccine, people, covid, vaccinated, day, t, year, thing, symptom, time
3	day, vaccine, pfizer, shot, vaccinated, dose, week, vaccination, good, pain
4	t, vaccine, people, covid, vaccinated, day, old, year, don, shot
5	vaccine, shot, day, time, people, covid, vaccinated, year, booster, week
6	vaccine, covid, people, long, mask, shot, symptom, t, vaccinated, day
7	vaccine, day, people, vaccinated, covid, dose, week, t, shot, vaccination
	21-Oct
0	vaccinated, day, vaccine, population, dose, shot, year, old, utahns, tot
1	vaccine, people, t, covid, effect, pfizer, day, shot, booster, vaccinated
2	vaccine, covid, shot, people, moderna, booster, t, pfizer, month, vaccinated
3	vaccine, day, covid, shot, week, effect, people, symptom, year, t
4	people, vaccine, day, shot, thing, vaccination, covid, booster, work, post
5	vaccine, covid, t, booster, fever, people, time, moderna, arm, shot
6	day, shot, covid, vaccine, t, pain, heart, time, week, hour
	shot, day, booster, vaccine, moderna, people, vaccinated, covid, good, pfizer
	21-Nov
0	shot, vaccine, people, day, thing, moderna, week, vaccinated, covid, booster
1	vaccine, people, shot, booster, covid, time, day, t, case, arm
2	booster, shot, vaccine, day, people, moderna, t, case, time, effect
3	vaccine, day, covid, people, booster, vaccinated, week, time, hour, shot
	vaccine, people, booster, covid, vaccinated, shot, day, vaccination, bad, month

Table 8 Continued

Topic	Words
5	people, vaccine, covid, day, pfizer, booster, shot, time, bad, vaccinated
6	vaccine, covid, vaccinated, t, booster, people, long, day, month, year
7	vaccine, vaccinated, shot, t, covid, people, year, day, time, effect
	21-Dec
0	|, datum, vaccine, day, people, covid, vaccinated, shot, case, booster
1	vaccine, booster, case, day, covid, people, datum, week, vaccinated, test
2	vaccine, covid, people, day, t, time, shot, booster, symptom, case
3	vaccine, booster, t, people, covid, day, vaccinated, shot, month, time
4	vaccine, covid, case, day, t, symptom, booster, shot, vaccination, week
5	vaccine, people, covid, shot, booster, vaccinated, omicron, t, day, time
6	vaccine, day, people, covid, month, pfizer, year, t, bad, booster
7	vaccine, booster, covid, t, week, day, people, time, vaccinated, shot
	Jan-22
0	vaccine, day, t, people, covid, shot, time, vaccinated, symptom, week
1	vaccine, covid, people, case, day, long, shot, time, t, symptom
2	|, case, day, people, vaccinated, booster, covid, t, vaccine, datum
3	booster, vaccine, day, covid, positive, t, time, bad, symptom, shot
4	day, vaccine, test, symptom, covid, positive, booster, pain, case, week
5	vaccine, people, day, covid, year, case, mask, test, thing, datum
6	vaccine, covid, day, booster, symptom, people, month, case, bad, t
7	vaccine, covid, day, people, vaccinated, t, booster, test, omicron, throat
	Feb-22
0	vaccine, people, day, covid, month, datum, booster, effect, vaccinated, shot
1	vaccine, covid, vaccinated, t, day, people, booster, vaccination, time, shot
2	covid, vaccine, people, long, t, day, year, shot, vaccinated, mask
3	people, vaccine, case, covid, vaccinated, day, test, datum, long, week
4	vaccine, people, datum, covid, case, vaccinated, booster, death, week, mask
5	vaccine, t, people, booster, time, mandate, day, shot, mask, week
6	vaccine, people, covid, t, shot, day, long, vaccinated, booster, week
7	vaccine, covid, day, people, case, t, vaccinated, week, booster, vaccination

Table 9: LDA Reddit 2.0 positive classified topics

Topic	Words
	20-Jan
0	study, foot, t, doctor, vaccine, people, effect, vaccinated, pain, thing
1	foot, t, effect, time, doctor, good, year, vaccination, muslim, pain
2	people, vaccine, t, thing, drug, vaccinated, death, case, doctor, need
3	doctor, vaccination, year, vaccine, virus, good, people, time, parent, lot
4	people, vaccine, disease, t, autoimmune, way, shot, need, thing, virus
5	sure, people, vaccinated, good, symptom, vaccine, t, effect, day, anti
6	vaccine, year, pain, t, foot, kid, disease, time, day, vaccinated
7	people, immune, system, good, virus, time, flu, child, vaccine, year
	20-Feb
0	case, vaccine, flu, virus, year, day, people, season, time, good
1	vaccine, virus, hospital, life, people, system, medic, immune, person, healthy
2	people, flu, case, pain, good, t, vitamin, immune, thing, day
3	virus, pain, people, week, patient, case, good, day, time, vaccine
4	t, people, time, vaccine, don, year, case, thing, kid, good
5	immune, system, time, people, day, flu, case, year, patient, sar
6	time, t, people, day, vaccine, work, bad, pain, thing, symptom
7	people, work, time, virus, case, flu, immune, day, year, vaccine
	20-Mar
0	people, virus, need, t, time, day, home, thing, case, work
1	people, day, case, time, virus, vaccine, immune, year, t, home
2	people, t, mask, need, good, virus, day, vaccine, thing, case
3	people, t, day, system, vaccine, immune, time, patient, home, work
4	people, need, day, virus, case, symptom, home, time, immune, bad
5	people, case, day, week, time, virus, coronavirus, test, t, mask
6	virus, people, good, time, immune, day, system, vaccine, thing, way
	t, people, case, need, virus, time, social, don, help, thing

Table 9 Continued

Topic	Words
	20-Apr
0	people, need, test, day, t, vaccine, time, thing, testing, week
1	people, need, test, day, home, time, testing, virus, t, work
2	people, vaccine, t, x, need, testing, home, new, thing, test
3	t, x, people, virus, need, vaccine, test, day, time, new
4	people, x, t, test, need, work, testing, case, state, day
5	x, t, people, need, day, vaccine, new, home, good, thing
6	people, t, new, test, x, state, testing, case, day, need
7	people, test, case, x, day, need, testing, t, death, work
	20-May
0	people, test, t, vaccine, case, number, antibody, testing, work, death
1	people, x, t, time, need, state, work, virus, testing, right
2	people, t, virus, vaccine, time, x, testing, day, thing, low
3	people, t, case, antibody, test, death, time, x, week, testing
4	people, t, day, time, x, case, good, testing, new, need
5	people, t, case, x, week, day, time, sure, virus, way
6	t, people, mask, need, day, thing, x, week, good, time
7	people, t, x, day, need, test, sure, state, case, work
	20-Jun
0	feme, case, county, t, , , disease, death, , x
1	case, x, t, test, people, cell, day, death, time, mask
2	day, case, county, t, people, feme, need, time, symptom, death
3	people, t, time, x, vaccine, case, week, thing, need, number
4	t, people, mask, test, death, virus, vaccine, x, good, work
5	people, vaccine, case, good, t, covid, day, effect, cell, immune
6	x, t, case, people, day, death, need, antibody, test, patient
7	people, t, virus, patient, thing, group, number, day, vaccine, study
	20-Jul
0	people, t, day, case, vaccine, mask, virus, time, thing, positive
1	day, case, test, vitamin, patient, people, infection, time, week, covi

Table 9 Continued

Topic	Words
2	case, people, county, t, day, feme, patient, test, time, mask
3	t, day, people, vaccine, time, x, mask, test, case, long
4	t, people, day, covid, don, vaccine, x, kid, good, test
5	vaccine, case, feme, county, t, people, hospit, death, patient, time
6	vaccine, people, day, t, x, hope, mask, immune, case, time
7	t, people, test, day, x, school, testing, time, need, student
	20-Aug
0	vaccine, people, case, vir, day, x, x, virus, time, child
1	t, mask, people, thing, case, x, work, time, day, good
2	case, test, day, people, vaccine, county, good, new, school, t
3	vaccine, day, virus, t, test, antibody, people, cell, response, thing
4	people, t, virus, immune, good, thing, time, day, need, system
5	t, people, vaccine, test, week, mask, x, day, case, thing
6	vaccine, day, t, people, time, case, thing, way, death, don
7	case, county, feme, disease, death, people, new, t, tot, test
	20-Sep
0	vaccine, people, mask, case, time, county, t, week, lot, right
1	day, t, vaccine, time, year, virus, people, thing, test, county
2	county, case, vitamin, patient, day, covid, feme, people, new, test
3	vaccine, people, week, thing, new, study, mask, good, covid, right
4	vaccine, test, family, good, patient, week, flu, people, covid, t
5	vitamin, patient, vaccine, people, calcifediol, mask, day, risk, icu, study
6	virus, vaccine, case, people, day, immunity, thing, mask, county, time
7	day, t, vaccine, vitamin, good, immune, thing, symptom, patient, year
	20-Oct
0	t, people, time, vaccine, cell, day, good, virus, covid, case
1	vaccine, system, immune, people, thing, virus, time, case, t, new
2	vaccine, people, t, immune, mask, time, risk, thing, case, county
3	t, thing, vaccine, immune, system, people, friend, good, covid, day

Table 9 Continued

Topic	Words
4	vaccine, people, year, good, t, time, flu, virus, immune, covid
5	people, time, case, mask, year, need, right, day, school, x
6	time, covid, t, mask, case, day, antibody, vaccine, need, people
7	day, people, patient, symptom, virus, mask, t, year, immune, case
	20-Nov
0	vaccine, covid, year, people, long, need, week, right, virus, flu
1	vaccine, year, time, covid, people, good, long, day, way, t
2	people, vaccine, time, covid, day, case, mask, t, work, year
3	day, people, vaccine, covid, week, case, test, new, year, t
4	vaccine, good, time, thing, right, people, need, day, sure, way
5	vaccine, t, people, year, day, time, case, right, new, news
6	vaccine, case, day, t, covid, people, test, time, good, positive
7	vaccine, case, covid, people, news, day, virus, t, year, high
	20-Dec
0	vaccine, people, day, t, covid, time, week, need, work, pfizer
1	vaccine, people, t, day, week, case, new, virus, hospit, care
2	vaccine, day, people, week, long, time, t, work, good, number
3	vaccine, day, covid, week, people, case, pfizer, work, t, time
4	vaccine, people, day, time, week, hope, year, datum, covid, new
5	vaccine, people, covid, time, day, week, t, thing, long, hospit
6	vaccine, good, people, coronavirus, question, day, covid, r, t, time
7	vaccine, people, case, long, t, day, good, dose, effect, month
	21-Jan
0	vaccine, day, t, people, time, long, week, covid, moderna, arm
1	vaccine, day, dose, people, shot, hour, arm, county, covid, case
2	day, vaccine, shot, week, people, covid, county, good, symptom, hour
3	vaccine, people, week, vaccination, day, county, covid, thing, time, year
4	vaccine, |, day, arm, sore, dose, county, good, week, people
5	vaccine, dose, day, people, effect, t, covid, shot, time, week
6	vaccine, shot, people, day, dose, covid, moderna, work, week, time

Table 9 Continued

Topic	Words
	21-Feb
0	shot, vaccine, covid, people, t, day, arm, vaccination, effect, bad
1	vaccine, day, dose, week, people, t, covid, county, tot, vaccination
2	day, vaccine, arm, people, shot, t, hour, week, time, moderna
3	vaccine, day, county, case, people, vaccination, effect, school, state, shot
4	vaccine, day, shot, t, people, covid, arm, sore, symptom, week
5	vaccine, day, dose, people, week, moderna, past, time, good, tot
6	|, vaccine, shot, hour, arm, time, week, good, symptom, pain
7	vaccine, day, people, dose, shot, county, week, covid, t, vaccinated
	21-Mar
0	vaccine, shot, day, arm, covid, people, t, effect, good, yesterday
1	day, vaccine, effect, week, shot, people, moderna, hour, symptom, dose
2	vaccine, day, covid, people, shot, week, symptom, arm, effect, vaccinated
3	shot, day, vaccine, t, appointment, today, time, case, week, bad
4	vaccine, people, shot, day, |, arm, dose, pfizer, covid, effect
5	day, vaccine, shot, people, vaccinated, t, dose, appointment, week, arm
6	vaccine, shot, day, bad, t, hour, arm, week, sore, year
7	vaccine, shot, day, t, people, week, arm, vaccination, time, vaccinated
	21-Apr
0	vaccine, shot, people, covid, time, t, effect, day, symptom, arm
1	vaccine, day, shot, pfizer, hour, covid, week, arm, t, period
2	vaccine, shot, day, people, effect, arm, vaccinated, covid, t, |
3	t, vaccine, day, covid, week, people, vaccinated, don, good, hour
4	vaccine, shot, day, case, t, people, time, bad, hour, week
5	day, shot, vaccine, week, people, effect, bad, symptom, t, pfizer
6	vaccine, shot, day, |, people, week, moderna, vaccinated, today, t
7	vaccine, effect, day, arm, t, thing, hour, site, bet, week
	21-May
0	vaccine, covid, shot, hour, day, week, symptom, arm, period, t
1	day, shot, vaccine, people, t, arm, effect, week, covid, Pfizer

Table 9 Continued

Topic	Words
2	vaccine, shot, day, symptom, week, t, body, covid, time, month
3	vaccine, day, people, vaccinated, shot, population, dose, year, old, covid
4	day, vaccine, shot, period, effect, vaccinated, people, year, time, t
5	|, vaccine, vaccinated, shot, week, day, pfizer, people, t, effect
6	shot, vaccine, effect, bad, people, day, week, covid, symptom, time
7	vaccine, week, shot, day, people, t, covid, good, month, vaccinated
	21-Jun
0	day, vaccine, shot, covid, week, effect, pfizer, people, vaccinated, symptom
1	day, vaccine, vaccinated, dose, shot, people, year, population, old, utahns
2	day, vaccine, vaccinated, dose, year, shot, population, period, people, time
3	vaccine, people, shot, vaccinated, day, effect, covid, time, pfizer, population
4	vaccine, people, t, shot, vaccinated, covid, day, symptom, thing, effect
5	vaccine, vaccinated, day, arm, t, week, pfizer, symptom, covid, sore
6	shot, day, pain, arm, week, vaccine, t, bad, people, effect
7	vaccine, shot, day, week, people, vaccination, t, effect, month, case
	21-Jul
0	shot, vaccine, day, vaccinated, good, dose, people, symptom, vaccination, t
1	vaccine, people, covid, vaccinated, shot, time, way, t, mask, risk
2	vaccine, day, week, people, covid, effect, shot, symptom, pain, month
3	vaccine, vaccinated, day, people, shot, dose, covid, pfizer, t, effect
4	vaccine, day, vaccinated, people, effect, shot, covid, time, week, t
5	vaccine, day, people, shot, arm, covid, pain, vaccinated, year, effect
6	day, pain, moderna, vaccine, pfizer, issue, arm, covid, doctor, shot
7	day, vaccine, t, people, vaccinated, covid, shot, pfizer, week, symptom
	21-Aug
0	vaccine, day, vaccinated, shot, t, people, year, covid, dose, symptom
1	vaccine, day, t, people, covid, vaccinated, bad, shot, good, symptom
2	vaccine, people, shot, arm, day, t, covid, effect, time, week
3	vaccine, covid, vaccinated, |, t, day, people, vaccination, year, case
4	day, vaccine, vaccinated, covid, week, effect, case, symptom, t, people

Table 9 Continued

Topic	Words
5	vaccine, shot, day, covid, time, norm, week, effect, t, vaccinated
6	vaccine, |, vaccinated, covid, day, people, dose, case, week, shot
7	day, shot, vaccinated, pfizer, vaccine, covid, people, moderna, year, good
	21-Sep
0	vaccine, covid, day, shot, people, t, bad, vaccinated, symptom, dose
1	shot, day, arm, vaccine, week, covid, effect, people, bad, time
2	vaccine, effect, people, shot, day, virus, covid, week, vaccinated, good
3	vaccine, people, vaccinated, booster, shot, covid, time, day, t, week
4	day, vaccinated, vaccine, dose, year, people, population, shot, old, tot
5	vaccine, people, week, covid, t, vaccinated, day, shot, time, symptom
6	vaccine, covid, people, day, vaccinated, month, shot, vaccination, mask, good
7	vaccine, covid, day, shot, vaccinated, t, people, death, high, effect
	21-Oct
0	shot, day, people, vaccinated, vaccine, moderna, t, week, pfizer, case
1	day, vaccine, shot, people, arm, pfizer, symptom, covid, effect, vaccinated
2	vaccine, day, people, good, t, shot, effect, work, bad, booster
3	vaccine, covid, t, shot, day, time, people, vaccinated, bad, hour
4	vaccine, shot, covid, booster, day, moderna, r, discussion, people, hour
5	vaccine, booster, shot, moderna, arm, t, day, case, vaccinated, hour
6	vaccine, covid, month, booster, week, effect, moderna, time, symptom, vaccinated
7	day, vaccine, vaccinated, dose, shot, population, year, people, tot, utahns
	21-Nov
0	booster, vaccinated, day, vaccine, shot, people, pfizer, old, year, effect
1	vaccine, shot, booster, day, covid, t, people, effect, vaccinated, moderna
2	vaccine, datum, covid, case, pa, shot, day, new, pennsylvania, topic
3	vaccine, covid, booster, people, vaccinated, time, day, good, year, month
4	vaccine, t, shot, booster, covid, time, day, moderna, pfizer, people
5	shot, vaccine, day, people, booster, covid, effect, vaccinated, hour, bad
6	vaccine, people, vaccinated, covid, day, year, shot, month, high, Pfizer

Table 9 Continued

Topic	Words
7	day, shot, booster, vaccine, week, arm, bad, effect, symptom, time
	21-Dec
0	day, booster, vaccine, covid, shot, moderna, t, bad, omicron, vaccinated
1	vaccine, covid, people, day, datum, vaccinated, case, dose, shot, symptom
2	booster, shot, covid, day, people, t, case, vaccine, vaccinated, test
3	vaccine, covid, time, t, people, shot, booster, bad, case, year
4	vaccine, case, day, booster, covid, people, new, good, datum
5	day, vaccine, case, people, covid, booster, vaccinated, new, datum, symptom
6	day, vaccine, covid, booster, t, shot, people, pfizer, week, long
7	vaccine, covid, people, omicron, day, vaccinated, t, virus, body, person
	Jan-22
0	vaccine, t, day, covid, booster, people, vaccinated, shot, case, omicron
1	day, booster, vaccine, shot, week, pain, bad, symptom, effect, time
2	covid, day, people, vaccine, case, bad, time, shot, vaccinated, symptom
3	vaccine, datum, t, day, covid, case, people, vaccinated, new, pa
4	day, covid, case, positive, symptom, booster, test, t, people
5	case, vaccine, people, week, day, new, t, positive, covid
6	day, booster, symptom, covid, vaccine, case, test, positive, t, throat
7	day, covid, case, week, time, booster, test, tot, positive
	Feb-22
0	vaccine, covid, datum, pa, long, day, case, pennsylvania, vaccinated, people
1	vaccine, case, datum, booster, t, shot, day, time, month, people
2	vaccine, covid, day, booster, shot, people, vaccinated, effect, t, moderna
3	vaccine, people, case, datum, vaccinated, day, covid, booster, t, death
4	people, vaccine, case, datum, day, vaccinated, covid, week, test, mask
5	people, day, vaccine, case, t, booster, week, time, covid, right
6	people, covid, vaccine, datum, case, long, time, shot, vaccinated, antibody
7	covid, vaccine, people, long, day, case, symptom, week, test, vaccinated

Table 10: LDA Twitter negative classified topics

Topic	Words
	20-Jan
0	immunology, vaccine, immunity, immune, industry, immunotherapy, doctor, rome, immune, scientist, heamatology, microbiology
1	bookmark, vaccine, immune, system, immunity, date, visit, vaccines, work, immunochemistry, conferences, immunotherapy, pathologic
2	presentation, poster, vaccine, rome, research, l, immunology, conference, opportunity
3	immunochemistry, immunology, virology, pathologic, immunotherapy, immune, chemistry, conferences, vaccineswork, visit, date, immunity
4	vaccine, immunology, rome, immunity, immunochemistry, pathology, immunotherapy, speaker
5	vaccine, immunology, immunity, rome, immune, chemistry, immunotherapy, pathology, virology
6	vaccine, rome, immunology, immunochemistry, food, safety, immunity, upcoming, foodscience
7	vaccine, immunology, rome, immunotherapy, immunity, immunochemistry, virology, expert
	20-Feb
0	vaccine, coronavirus, month, ready, trump, covid, new, flu, cdc
1	vaccine, coronavirus, covid, ready, new, month, virus, sar, flu, covid
2	vaccine, coronavirus, flu, covid, virus, trials, drug, covid, good, year
3	vaccine, covid, people, company, drug, sar, time, moderna, flu, batch
4	vaccine, flu, virus, covid, year, coronavirus, treatment, people, available
5	vaccine, coronavirus, covid, drug, virus, year, treatment, people, need, amp
6	vaccine, coronavirus, covid, flu, prevent, death, pandemic, china, disease
7	vaccine, flu, people, rate, virus, coronavirus, covid, month, new
	20-Mar
0	vaccine, coronavirus, covid, virus, human, time, covid, test, trial, flu
1	vaccine, flu, virus, covid, people, need, research, world, rate
2	vaccine, covid, virus, treatment, research, rate, johnson, coronavirus, good

Table 10 Continued

Topic	Words
3	vaccine, covid, flu, coronavirus, people, treatment, time, death, way, cure
4	vaccine, covid, people, flu, virus, coronavirus, world, immunity, company
5	vaccine, covid, people, flu, virus, year, trump, coronavirus, time
6	vaccine, covid, flu, virus, coronavirus, year, prevent, disease, need, rate
7	vaccine, covid, flu, virus, test, new, month, good, people
	20-Apr
0	vaccine, covid, people, virus, test, flu, need, time, world
1	vaccine, covid, world, coronavirus, year, new, gate, pandemic, death
2	vaccine, covid, flu, coronavirus, people, need, virus, testing, immunity, death
3	vaccine, coronavirus, covid, covid, trials, people, scientist, test, human, flu
4	vaccine, covid, flu, people, death, testing, year, virus, africa, treatment
5	vaccine, covid, work, people, treatment, year, flu, virus, trump, amp
6	vaccine, covid, virus, gate, people, human, flu, need, trial
7	vaccine, covid, virus, people, treatment, cure, time, world, research
	20-May
0	vaccine, covid, china, test, coronavirus, need, covid, trump, people
1	vaccine, covid, coronavirus, trump, world, death, pandemic, new, people
2	vaccine, covid, flu, coronavirus, virus, covid, people, long, research
3	vaccine, covid, flu, people, year, virus, treatment, need, gate
4	vaccine, covid, world, virus, coronavirus, people, news, death, time
5	vaccine, covid, people, world, virus, time, trump, good, death
6	vaccine, covid, flu, trials, clinic, death, trial, trump, china, today
7	vaccine, covid, coronavirus, virus, new, work, people, year, trump, trials
	20-Jun
0	vaccine, covid, people, year, virus, world, trial, development, cure
1	vaccine, covid, flu, year, people, available, need, good, end
2	vaccine, covid, people, flu, virus, time, need, new, china, amp
3	vaccine, covid, virus, research, scientist, flu, year, need, public
4	vaccine, covid, coronavirus, people, news, time, trial, spread, virus
5	vaccine, covid, virus, trials, people, good, gate, coronavirus, treatment

Table 10 Continued

Topic	Words
6	vaccine, covid, people, coronavirus, world, trump, testing, mask, end
7	vaccine, covid, year, trials, news, mask, covid, world, clinic, amp
	20-Jul
0	vaccine, covid, trial, news, good, oxford, phase, need, available, trials
1	vaccine, covid, flu, mask, people, long, way, work, time
2	vaccine, covid, new, coronavirus, news, covid, year, death, oxford
3	vaccine, covid, people, immunity, human, coronavirus, need, work, trials, testing
4	vaccine, covid, trump, covid, trials, mark, news, china, child
5	vaccine, covid, flu, trials, world, news, virus, year, people, good
6	vaccine, covid, people, virus, year, russia, flu, death, coronavirus
7	vaccine, covid, people, flu, research, trial, need, coronavirus, india, world
	20-Aug
0	vaccine, covid, russia, world, virus, trial, putin, need, year
1	vaccine, covid, russia, flu, people, virus, work, life, thing
2	vaccine, covid, flu, year, new, need, trump, covidvaccine, immunity, news
3	vaccine, covid, flu, russia, effective, news, coronavirus, country, covid
4	vaccine, covid, flu, people, trump, russia, month, new, covid, mask
5	vaccine, covid, people, russia, flu, world, year, trials, putin, coronavirus
6	vaccine, covid, trump, russia, good, trials, people, flu, virus
7	vaccine, covid, trump, india, russia, people, year, world, country, amp
	20-Sep
0	vaccine, covid, flu, people, trump, year, need, available, mask
1	vaccine, covid, people, flu, trial, mask, trump, trials, country, death
2	vaccine, covid, trump, people, news, cdc, trial, state, trials
3	vaccine, covid, trump, astrazeneca, people, good, trial, virus, study
4	vaccine, covid, trump, cdc, case, year, trials, people, end
5	vaccine, covid, trump, news, coronavirus, |, election, people, trials, time
6	vaccine, covid, flu, people, year, virus, trials, shot, time, amp
7	vaccine, covid, trump, people, end, work, safe, election, trials
	20-Oct

Table 10 Continued

Topic	Words
0	vaccine, covid, trump, people, flu, free, need, trial, new
1	vaccine, covid, trump, people, virus, flu, safe, year, time
2	vaccine, covid, johnson, people, plan, illness, new, company, year
3	vaccine, covid, year, trump, people, trials, death, plan, johnson
4	vaccine, flu, covid, year, virus, shot, people, trial, need
5	vaccine, covid, trial, people, immunity, johnson, mask, trials, coronavirus
6	vaccine, covid, people, trump, plan, world, news, pfizer, free, day
7	vaccine, covid, new, news, government, country, india, effective, people
	20-Nov
0	vaccine, covid, pfizer, news, trump, effective, people, election, world
1	vaccine, covid, effective, pfizer, people, new, year, coronavirus, flu, news
2	vaccine, covid, pfizer, people, news, biden, effective, flu, year
3	vaccine, covid, pfizer, covidvaccine, effective, datum, need, trump, result, company
4	vaccine, covid, people, month, virus, good, covidvaccine, flu, day
5	vaccine, covid, news, pfizer, people, effective, covid, covidvaccine, trump
6	vaccine, covid, pfizer, trump, effective, biden, biontech, trial, president, people
7	vaccine, covid, people, day, pfizer, flu, trump, effective, time, thing
	20-Dec
0	vaccine, covid, people, need, covidvaccine, new, flu, virus, long
1	vaccine, covid, people, trump, time, worker, good, pfizer, day, news
2	vaccine, covid, covidvaccine, covid, people, pfizer, news, state, world
3	vaccine, covid, people, need, covidvaccine, black, country, death, government
4	vaccine, covid, trump, dose, people, covidvaccine, pfizer, uk, week, news
5	vaccine, covid, year, people, covidvaccine, work, safe, covid, risk
6	vaccine, covid, covidvaccine, trump, year, pfizer, today, right, covid
7	vaccine, covid, pfizer, covidvaccine, people, effect, covid, flu, use
	21-Jan
0	vaccine, covid, people, need, flu, worker, new, country, year
1	vaccine, covid, year, old, covidvaccine, need, new, people, good, free

Table 10 Continued

Topic	Words
2	vaccine, covid, people, death, effect, day, time, worker, risk
3	vaccine, covid, covidvaccine, year, virus, work, new, people, dose
4	vaccine, covid, covidvaccine, new, county, people, today, time, news, amp
5	vaccine, covid, people, new, pfizer, dose, day, covidvaccine, vaccination, year
6	vaccine, covid, covidvaccine, people, covid, vaccination, india, worker, care
7	vaccine, covid, covidvaccine, time, today, day, |, need, covid
	21-Feb
0	vaccine, covid, today, state, week, shot, people, india, year
1	vaccine, covid, people, good, today, covidvaccine, death, new, year
2	vaccine, covid, flu, year, covidvaccine, people, virus, death, time
3	vaccine, covid, people, covidvaccine, new, covid, variant, pfizer, country
4	vaccine, covid, dose, people, need, day, work, covidvaccine, vaccination
5	vaccine, covid, people, johnson, appointment, covidvaccine, county, week, time
6	vaccine, covid, covidvaccine, people, vaccination, covid, day, need, dose
7	vaccine, covid, people, effect, effective, long, risk, year, old, new
	21-Mar
0	vaccine, covid, people, country, day, vaccinated, mask, india, vaccination
1	vaccine, covid, people, shot, today, dose, new, appointment, week
2	vaccine, covid, week, people, time, risk, vaccinated, age, eligible
3	vaccine, covid, johnson, covidvaccine, vaccination, mask, long, school, free
4	vaccine, covid, people, covidvaccine, covid, old, safe, |, eligible
5	vaccine, covidvaccine, covid, trump, covid, worker, vaccinated, people, great
6	vaccine, covid, people, death, effect, trump, covidvaccine, vaccinated, country
7	vaccine, covid, people, today, day, year, vaccination, news, good
	21-Apr
0	vaccine, covid, covidvaccine, people, need, new, case, india, vaccinated, week
1	vaccine, covid, passport, country, day, vaccination, people, virus, time
2	vaccine, covid, effect, people, india, dose, case, risk, passport, day
3	vaccine, covid, covidvaccine, long, people, vaccinated, appointment, time, covid
4	vaccine, covid, people, risk, death, vaccinated, year, mask, virus, need

Table 10 Continued

Topic	Words
5	vaccine, covid, shot, pfizer, today, people, moderna, covidvaccine, work, year
6	vaccine, covid, johnson, today, year, covidvaccine, blood, week, clot
7	vaccine, covid, people, vaccination, appointment, shot, vaccinated, day, covidvaccine
	21-May
0	vaccine, covid, people, vaccinated, good, mask, covidvaccine, vaccination, death
1	vaccine, covid, people, vaccination, today, year, covidvaccine, time, variant
2	vaccine, covid, people, need, child, shot, news, country, new
3	vaccine, covid, shot, day, year, people, covidvaccine, india, time, case
4	vaccine, covid, people, risk, need, effect, vaccinated, long, india
5	vaccine, covid, pfizer, death, india, vaccinated, dose, vaccination, covidvaccine
6	vaccine, covid, people, vaccinated, time, dose, vaccination, death, long
7	vaccine, covid, covidvaccine, shot, pfizer, day, covid, moderna, appointment, people
	21-Jun
0	vaccine, covid, people, need, death, vaccination, long, day, shot
1	vaccine, covid, people, vaccination, thing, life, good, country, vaccinated
2	vaccine, covid, people, vaccinated, year, death, covidvaccine, india, virus, case
3	vaccine, covid, people, risk, dose, child, covidvaccine, vaccinated, shot
4	vaccine, covid, covidvaccine, dose, shot, people, vaccination, available, age
5	vaccine, covid, people, vaccination, covidvaccine, covid, dose, today, vaccinated
6	vaccine, covid, government, people, free, jab, vaccinated, covidvaccine, country
7	vaccine, covid, people, vaccinated, variant, covidvaccine, hospit, vaccination, work
	21-Jul
0	vaccine, covid, people, vaccinated, variant, covidvaccine, unvaccinated, death, mask
1	vaccine, covid, people, vaccinated, free, life, covidvaccine, long, death
2	vaccine, covid, people, death, risk, long, vaccinated, vaccination, year
3	vaccine, covid, people, shot, variant, vaccinated, year, day, pfizer

Table 10 Continued

Topic	Words
4	vaccine, covid, vaccinated, risk, covidvaccine, case, people, system, pfizer, day
5	vaccine, covid, people, vaccinated, mask, work, case, government, news, pfizer
6	vaccine, covid, new, vaccination, people, death, vaccinated, state, help, amp
7	vaccine, covid, flu, good, vaccinated, virus, need, variant, long, amp
	21-Aug
0	vaccine, covid, people, vaccinated, risk, virus, vaccination, long, life, immune
1	vaccine, covid, people, effect, long, bad, time, term, pfizer
2	vaccine, covid, people, vaccinated, mask, sick, safe, mandate, hospit
3	vaccine, covid, mask, covidvaccine, anti, people, shot, school, work
4	vaccine, covid, vaccinated, people, shot, year, flu, case, risk
5	vaccine, covid, people, vaccinated, mask, variant, work, need, new
6	vaccine, covid, people, death, virus, datum, need, world, vaccinated, rate
7	vaccine, covid, covidvaccine, people, anti, mask, test, vaccination, vaccinated, state
	21-Sep
0	vaccine, covid, vaccinated, people, vaccination, covidvaccine, covid, way, child
1	vaccine, covid, people, vaccinated, mask, death, case, vaccination, risk, mandate
2	vaccine, covid, people, death, vaccinated, long, virus, mask, need
3	vaccine, covid, people, vaccinated, year, pfizer, free, need, good
4	vaccine, covid, people, right, variant, immunity, new, vaccinated, effective, good
5	vaccine, covid, people, child, year, anti, death, person, fact
6	vaccine, covid, people, risk, time, vaccinated, effect, long, thing, hospit
7	vaccine, covid, test, vaccinated, people, biden, time, shot, flu, amp
	21-Oct
0	vaccine, covid, people, month, vaccinated, need, death, shot, year
1	vaccine, covid, people, child, vaccinated, pfizer, vaccination, science, medic, risk
2	vaccine, covid, shot, people, flu, vaccinated, year, mask, mandate
3	vaccine, covid, mandate, kid, risk, people, immunity, vaccination, vaccinated, death
4	vaccine, covid, people, vaccinated, death, time, mandate, news, need, booster
5	vaccine, covid, death, vaccinated, virus, country, work, high, pandemic, new

Table 10 Continued

Topic	Words
6	vaccine, covid, people, flu, vaccinated, year, death, work, risk
7	vaccine, covid, people, vaccinated, effect, long, death, shot, immunity, amp
	21-Nov
0	vaccine, covid, people, mandate, booster, pfizer, year, kid, pandemic
1	vaccine, covid, people, pfizer, vaccinated, child, good, booster, work
2	vaccine, covid, death, vaccinated, year, flu, effect, case, shot, life
3	vaccine, covid, kid, risk, people, shot, booster, vaccinated, child
4	vaccine, covid, kid, child, year, age, long, shot, parent, school
5	vaccine, covid, people, vaccinated, year, kid, death, day, shot
6	vaccine, covid, people, death, work, month, vaccinated, immunity, risk
7	vaccine, covid, child, mandate, people, vaccinated, pfizer, age, virus, cdc
	21-Dec
0	vaccine, covid, people, omicron, risk, booster, variant, vaccinated, case, death
1	vaccine, covid, covidvaccine, vaccination, covid, booster, patient, omicron, people, today
2	vaccine, covid, people, year, virus, immunity, hospit, work, pfizer
3	vaccine, covid, people, vaccinated, death, mandate, mask, risk, shot
4	vaccine, covid, booster, day, flu, year, shot, new, anti, mandate
5	vaccine, covid, people, virus, death, time, booster, new, mandate
6	vaccine, covid, people, work, vaccinated, booster, long, year, variant, flu
7	vaccine, covid, people, year, covidvaccine, booster, vaccinated, child, world, lamb
	Jan-22
0	vaccine, covid, people, year, mandate, booster, virus, country, effective
1	covid, vaccine, people, vaccinated, work, year, time, mask, death, prevent
2	vaccine, covid, death, mrna, life, new, booster, question, mandate, vaccination
3	vaccine, covid, booster, people, mandate, need, child, school, kid, day
4	vaccine, covid, death, case, time, people, long, datum, sick
5	vaccine, covid, people, vaccinated, risk, work, bad, long, new
6	vaccine, covid, vaccinated, omicron, good, virus, year, people, time, mask
7	vaccine, covid, vaccinated, virus, booster, child, people, pfizer, vaccination

Table 10 Continued

Topic	Words
	Feb-22
0	vaccine, covid, people, case, mandate, vaccinated, death, right, risk
1	vaccine, covid, people, risk, death, vaccinated, thing, day, way, good
2	vaccine, covid, people, mandate, kid, vaccinated, long, flu, need, booster
3	vaccine, covid, people, year, death, vaccinated, mask, child, time
4	vaccine, covid, people, work, risk, mandate, year, government, datum, vaccination
5	vaccine, covid, people, country, child, risk, long, death, mask
6	vaccine, covid, people, work, death, long, mask, vaccinated, virus
7	vaccine, covid, booster, time, people, state, today, death, vaccinated

Table 11: LDA Twitter positive classified topics

Topic	Words
	20-Jan
0	poster, presentation, opportunity, l, research, immunology, conference, inflammation,
1	vaccine, immunology, rome, immunity, immunotherapy, immunochemistry, immune
2	immunochemistry, vaccines, work, date
3	immunology, vaccine, rome, immunochemistry, immunity, immunotherapy, poster,
4	vaccine, ity, foodsafy, rome, immunity, upcoming, expert, pathology, immunology
5	vaccine, immunology, rome, immunity, immunotherapy, immunochemistry, poster
6	vaccine, immunology, Italy, rome, immunochemistry, immunity, immunotherapy
7	immunology, vaccine, rome, immunotherapy, book, scientific, researcher, heamatology,
	20-Feb
0	vaccine, coronavirus, flu, virus, ready, prevent, month, available, moderna, year
1	vaccine, coronavirus, covid, people, month, time, virus, flu, need
2	vaccine, coronavirus, good, covid, news, virus, new, people, |
3	vaccine, coronavirus, development, drug, |, trials, new, human, lab, time
4	vaccine, covid, virus, flu, development, new, time, treatment, people
5	vaccine, coronavirus, novel, development, advance, drug, trials, novavax, covid, news
6	vaccine, drug, coronavirus, virus, trials, month, trump, flu, risk, covid
7	vaccine, covid, coronavirus, flu, new, research, ready, year, trials, month
	20-Mar
0	vaccine, covid, coronavirus, test, virus, treatment, flu, people, world, cure
1	vaccine, people, coronavirus, covid, hope, virus, time, development, scientist
2	vaccine, covid, coronavirus, flu, year, people, month, good, treatment
3	vaccine, coronavirus, covid, virus, covid, treatment, research, clinic, flu, news
4	vaccine, coronavirus, virus, prevent, good, china, people, covid, university, way
5	vaccine, covid, new, available, people, world, time, trials, month, pneumonia
6	vaccine, flu, covid, virus, work, research, good, help, new
7	vaccine, covid, treatment, coronavirus, need, people, research, available, cure
	20-Apr
0	vaccine, covid, coronavirus, help, scientist, covid, testing, new, research, time
1	vaccine, covid, need, virus, world, scientist, treatment, development, covid, amp

Table 11 Continued

Topic	Words
2	vaccine, covid, treatment, development, new, coronavirus, test, year, virus, covid
3	vaccine, covid, coronavirus, world, treatment, india, test, virus, cure
4	vaccine, covid, development, work, clinic, world, immunity, trials, virus, coronavirus
5	vaccine, covid, good, trials, pandemic, flu, human, people, way
6	vaccine, covid, people, virus, test, time, year, need, covid
7	vaccine, covid, coronavirus, need, people, new, news, flu, treatment, drug
	20-May
0	vaccine, covid, trial, treatment, need, people, coronavirus, hope, development, clinic
1	vaccine, covid, work, world, race, glob, available, antibody, treatment
2	vaccine, covid, people, world, trials, year, virus, flu, potenti, time
3	vaccine, covid, coronavirus, virus, available, time, covid, trial, development
4	vaccine, covid, people, year, flu, pandemic, virus, life, good
5	vaccine, covid, covid, coronavirus, treatment, news, development, research, good
6	vaccine, covid, people, death, end, need, coronavirus, good, flu, disease
7	vaccine, covid, coronavirus, need, news, world, development, trial, human, people
	20-Jun
0	vaccine, covid, treatment, news, virus, future, time, world, test, amp
1	vaccine, covid, coronavirus, phase, world, year, covid, development, new, amp
2	vaccine, covid, people, trials, trial, astrazeneca, human, potenti, covid
3	vaccine, covid, treatment, world, trials, clinic, trial, people, work
4	vaccine, covid, available, end, year, trials, news, people, development
5	vaccine, covid, need, people, good, available, case, help, right
6	vaccine, covid, flu, year, covid, virus, people, treatment, new
7	vaccine, covid, people, safe, good, need, free, death, india
	20-Jul
0	vaccine, covid, need, mask, people, india, human, trials, trial
1	vaccine, covid, news, people, oxford, uk, good, university, trials
2	vaccine, covid, trials, clinic, russia, news, coronavirus, virus, human
3	vaccine, covid, trials, covid, news, phase, trial, result, world, hope
4	vaccine, covid, trial, people, immune, need, response, year, safe, news

Table 11 Continued

Topic	Words
5	vaccine, covid, trials, human, new, covid, year, need, news, immunity
6	vaccine, covid, oxford, trials, people, news, coronavirus, human, good, time
7	vaccine, covid, trial, flu, india, trials, year, phase, coronavirus
	20-Aug
0	vaccine, covid, year, covid, coronavirus, good, trials, flu, trial, china
1	vaccine, covid, russia, available, world, trials, time, good, need, effective
2	vaccine, russia, covidvaccine, covid, world, putin, trial, president, russian, coronavirus
3	vaccine, covid, russia, world, people, trials, news, putin, flu, country
4	vaccine, covid, russia, news, world, trump, people, putin, |
5	vaccine, covid, russia, india, phase, safe, trials, world, clinic, dose
6	vaccine, covid, russia, research, time, safe, people, flu, end
7	vaccine, covid, people, russia, flu, india, putin, country, good
	20-Sep
0	vaccine, covid, safe, covid, trump, development, coronavirus, trial, effective
1	vaccine, covid, news, flu, virus, good, shot, work, death
2	vaccine, covid, trials, johnson, trial, |, covid, safe, astrazeneca
3	vaccine, covid, phase, trials, trial, covid, clinic, safe, month, candidate
4	vaccine, covid, new, flu, trials, available, year, trump, people
5	vaccine, covid, people, flu, trump, year, need, mask, trial
6	vaccine, covid, flu, trials, work, risk, clinic, trial, distribution
7	vaccine, covid, safe, trump, people, effective, need, time, flu
	20-Oct
0	vaccine, covid, trump, immunity, plan, india, herd, johnson, people
1	vaccine, covid, trump, hope, country, new, plan, china, treatment, government
2	vaccine, covid, people, flu, trump, death, year, time, president
3	vaccine, covid, flu, year, people, free, shot, need, trump
4	vaccine, covid, new, covid, good, trials, effective, time, safe, people
5	vaccine, covid, free, people, trump, virus, covid, available, safe
6	vaccine, covid, need, free, help, world, state, people, way, flu
7	vaccine, covid, year, news, available, trump, trial, trials, flu

Table 11 Continued

Topic	Words
	20-Nov
0	vaccine, covid, news, pfizer, trump, |, moderna, week, long, result
1	vaccine, covid, effective, pfizer, news, trump, moderna, people, coronavirus, good
2	vaccine, covid, effective, covidvaccine, pfizer, time, trump, dose, covid
3	vaccine, covid, people, news, biden, flu, trump, pfizer, year
4	vaccine, covid, pfizer, effective, biontech, trial, news, people, coronavirus, month
5	vaccine, covid, pfizer, news, effective, day, covidvaccine, trial, result, new
6	vaccine, covid, effective, news, pfizer, covidvaccine, people, moderna, good, trial
7	vaccine, covid, pfizer, effective, news, year, world, candidate, time, people
	20-Dec
0	vaccine, covid, news, new, uk, pfizer, covid, people, trump, world
1	vaccine, covid, pfizer, covidvaccine, uk, year, covid, people, world
2	vaccine, covid, covidvaccine, covid, pfizer, dose, need, week, public
3	vaccine, covid, pfizer, risk, biontech, year, people, president, use, public
4	vaccine, covid, people, pfizer, work, covidvaccine, need, month, available
5	vaccine, covid, covidvaccine, week, people, year, day, worry, need, patient
6	vaccine, covid, pfizer, today, week, worker, covidvaccine, thank, long, need
7	vaccine, covid, people, covidvaccine, today, pfizer, good, day, time
	21-Jan
0	vaccine, covid, covidvaccine, people, vaccination, need, today, new, effect, week
1	vaccine, covid, people, today, covidvaccine, covid, india, year, day, amp
2	vaccine, covid, covidvaccine, year, people, new, thank, virus, risk
3	vaccine, covid, need, new, covidvaccine, people, use, time, week
4	vaccine, covid, today, good, day, people, new, work, year
5	vaccine, covid, covidvaccine, need, today, india, work, good, shot, week
6	vaccine, covid, covidvaccine, people, day, time, worker, vaccination, week, need
7	vaccine, covid, state, vaccination, available, covidvaccine, today, dose, people
	21-Feb
0	vaccine, covid, dose, covidvaccine, week, resident, people, old, clinic, amp
1	vaccine, covid, dose, today, day, help, news, good, covidvaccine, pfizer

Table 11 Continued

Topic	Words
2	vaccine, covid, people, today, day, county, need, shot, death
3	vaccine, covidvaccine, covid, today, shot, trial, virus, week, question
4	vaccine, covidvaccine, covid, vaccination, year, need, number, people, today
5	vaccine, covid, today, covidvaccine, vaccination, appointment, week, covid, people
6	vaccine, covid, good, new, need, people, year, covidvaccine, news
7	vaccine, covid, covidvaccine, people, vaccination, vaccinated, today, community, free
	21-Mar
0	vaccine, covid, people, covidvaccine, appointment, vaccination, week, today, good
1	vaccine, covid, covidvaccine, today, india, old, people, vaccinated, year
2	vaccine, covid, today, covidvaccine, covid, shot, year, time, death
3	vaccine, covid, people, covidvaccine, vaccinated, vaccination, covid, eligible, week
4	vaccine, covid, year, eligible, today, covidvaccine, people, variant, march, country
5	vaccine, covid, people, covidvaccine, appointment, shot, thank, country, day
6	vaccine, covid, day, covidvaccine, effect, long, dose, appointment, new, week
7	vaccine, covid, johnson, today, shot, covidvaccine, need, help, vaccinated
	21-Apr
0	vaccine, covid, shot, covidvaccine, today, new, moderna, covid, vaccination
1	vaccine, covid, people, today, need, year, dose, risk, eligible
2	vaccine, covid, people, vaccinated, covidvaccine, covid, help, shot, good
3	vaccine, covid, covidvaccine, vaccinated, day, shot, people, week, appointment, today
4	vaccine, covid, appointment, day, today, eligible, week, covidvaccine, people, way
5	vaccine, covid, covidvaccine, people, good, today, pfizer, moderna, day, amp
6	vaccine, covid, appointment, vaccination, today, covidvaccine, available, people, shot,
7	vaccine, covid, shot, johnson, vaccination, appointment, vaccinated, work, long
	21-May
0	vaccine, covid, vaccinated, people, need, effect, appointment, shot, good
1	vaccine, covid, people, country, dose, covidvaccine, vaccinated, india, vaccination
2	vaccine, covid, people, vaccination, today, vaccinated, covidvaccine, day, india
3	vaccine, covid, india, pfizer, vaccination, people, year, covidvaccine, week
4	vaccine, covid, covidvaccine, vaccinated, appointment, pfizer, vaccination, covid, help,

Table 11 Continued

Topic	Words
5	vaccine, covid, vaccinated, people, death, mask, time, free, new, today
6	vaccine, covid, shot, covidvaccine, today, good, risk, people, pfizer
7	vaccine, covid, day, vaccinated, year, people, covidvaccine, pfizer, appointment, death
	21-Jun
0	vaccine, covid, covidvaccine, dose, safe, effective, need, case, vaccination
1	vaccine, covid, people, covidvaccine, time, vaccination, vaccinated, shot, new, dose
2	vaccine, covid, people, year, pfizer, clinic, need, today, hospit, shot
3	vaccine, covid, vaccinated, people, today, shot, covidvaccine, long, free
4	vaccine, covid, people, vaccinated, good, free, country, time, work
5	vaccine, covid, covidvaccine, vaccination, today, day, week, dose, people
6	vaccine, covid, people, vaccinated, risk, effect, vaccination, day, covid
7	vaccine, covid, covidvaccine, vaccination, death, vaccinated, free, today, people, dose
	21-Jul
0	vaccine, covid, today, people, vaccination, covid, case, immunity, dose, need
1	vaccine, covid, vaccinated, risk, people, vaccination, long, pfizer, covidvaccine
2	vaccine, covid, people, new, covidvaccine, vaccinated, variant, dose, week
3	vaccine, covid, people, death, dose, vaccination, good, need, safe
4	vaccine, covid, people, death, effect, variant, time, shot, work
5	vaccine, covid, people, vaccinated, death, vaccination, risk, life, shot
6	vaccine, covid, vaccinated, people, pfizer, long, variant, today, good, year
7	covid, vaccine, case, vaccinated, covidvaccine, day, vaccination, mask, people, jab
	21-Aug
0	vaccine, covid, vaccinated, vaccination, death, people, covidvaccine, good, variant
1	vaccine, covid, vaccinated, people, case, shot, vaccination, free, covidvaccine, today
2	vaccine, covid, people, shot, pfizer, mask, death, risk, need
3	vaccine, covid, people, covidvaccine, vaccination, year, day, vaccinated, new, long
4	vaccine, covid, people, vaccinated, death, good, effect, day, long, unvaccinated
5	vaccine, covid, vaccinated, people, covidvaccine, spread, vaccination, mask, death
6	vaccine, covid, people, risk, vaccinated, child, effect, dose, safe
7	vaccine, covid, people, vaccinated, year, anti, life, long, mask, death

Table 11 Continued

Topic	Words
	Sep-21
0	vaccine, covid, vaccination, need, mask, help, people, death, time
1	vaccine, covid, people, vaccinated, kid, shot, need, death, work
2	vaccine, covid, people, vaccinated, death, year, work, safe, good
3	vaccine, covid, vaccinated, people, risk, long, shot, child, time
4	vaccine, covid, people, vaccinated, hospit, life, chance, point, issue
5	vaccine, covid, vaccinated, people, mask, death, unvaccinated, today, likely, risk
6	vaccine, covid, death, year, people, effective, risk, vaccinated, infection, flu
7	vaccine, covid, people, virus, immunity, need, vaccination, mask, effective, amp
	21-Oct
0	vaccine, covid, people, work, vaccinated, good, available, system, case, infection
1	vaccine, covid, people, death, vaccinated, effect, kid, new, work, booster
2	vaccine, covid, people, vaccinated, good, anti, vaccination, pfizer, unvaccinated
3	vaccine, covid, flu, booster, vaccination, school, risk, death, vaccinated, people
4	vaccine, covid, booster, people, case, vaccinated, year, need, available
5	vaccine, covid, people, vaccinated, shot, flu, year, death, risk
6	vaccine, covid, vaccinated, people, immunity, child, work, shot, day
7	vaccine, covid, people, risk, safe, vaccinated, death, time, effective, child
	21-Nov
0	vaccine, covid, booster, pfizer, shot, people, child, today, age, kid
1	vaccine, covid, people, kid, child, long, year, vaccinated, risk
2	vaccine, covid, risk, booster, kid, flu, age, available, people, child
3	vaccine, covid, people, death, year, effect, old, long, clinic
4	vaccine, covid, death, kid, shot, vaccinated, today, work, time, rate
5	vaccine, covid, time, year, people, work, immunity, child, booster, new
6	vaccine, covid, vaccinated, year, child, vaccination, people, booster, death, today
7	vaccine, covid, vaccinated, people, child, booster, shot, covidvaccine, covid
	21-Dec
0	vaccine, covid, good, omicron, booster, new, news, shot, case, child
1	vaccine, covid, booster, covidvaccine, omicron, work, pfizer, covid, year, dose

Table 11 Continued

Topic	Words
2	vaccine, covid, people, booster, work, shot, year, death, vaccinated
3	vaccine, covid, booster, people, need, dose, risk, vaccination, case, pfizer
4	vaccine, covid, people, booster, death, vaccinated, need, child, risk, high
5	vaccine, covid, year, people, vaccination, vaccinated, death, booster, omicron
6	vaccine, covid, booster, people, vaccinated, shot, flu, year, good
7	vaccine, covid, people, vaccinated, booster, time, shot, flu, death
	Jan-22
0	vaccine, covid, vaccinated, death, work, people, variant, case, omicron
1	vaccine, covid, booster, covidvaccine, shot, people, today, clinic, vaccination
2	vaccine, covid, people, vaccinated, risk, booster, work, day, infection
3	vaccine, covid, death, year, people, mask, kid, booster, vaccination
4	vaccine, covid, risk, death, people, vaccination, bad, need, child
5	vaccine, covid, people, vaccinated, year, time, mask, new, long, thing
6	vaccine, covid, vaccinated, booster, pfizer, life, child, people, year
7	vaccine, covid, booster, vaccinated, people, long, available, death, child, year
	Feb-22
0	vaccine, covid, people, booster, vaccinated, year, work, child, need
1	vaccine, covid, vaccinated, child, long, year, effective, work, vaccination
2	vaccine, covid, need, new, child, pfizer, covidvaccine, year, people
3	vaccine, covid, people, death, long, vaccinated, booster, vaccination, need, right
4	vaccine, covid, vaccinated, people, death, risk, effective, prevent, hospit
5	vaccine, covid, booster, people, death, time, risk, vaccinated, effect, shot
6	vaccine, covid, child, vaccination, mask, risk, vaccinated, virus, safe, pandemic
7	vaccine, covid, people, risk, year, mask, kid, vaccinated, dose, long

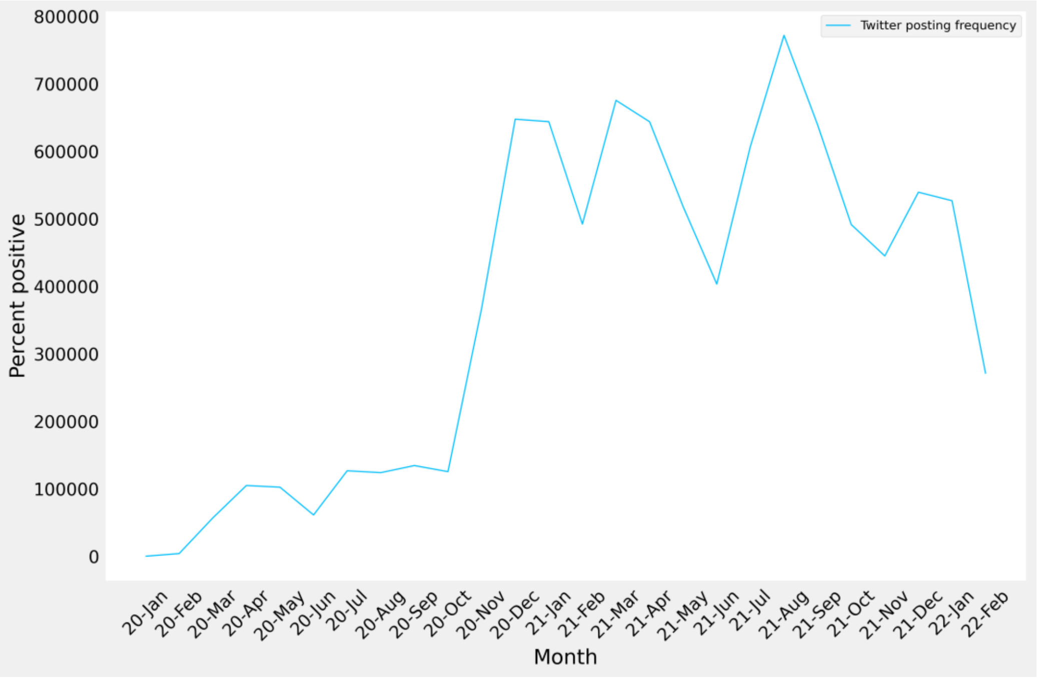


Figure 1: Twitter posting frequency from January 1, 2020 to March 1, 2022.

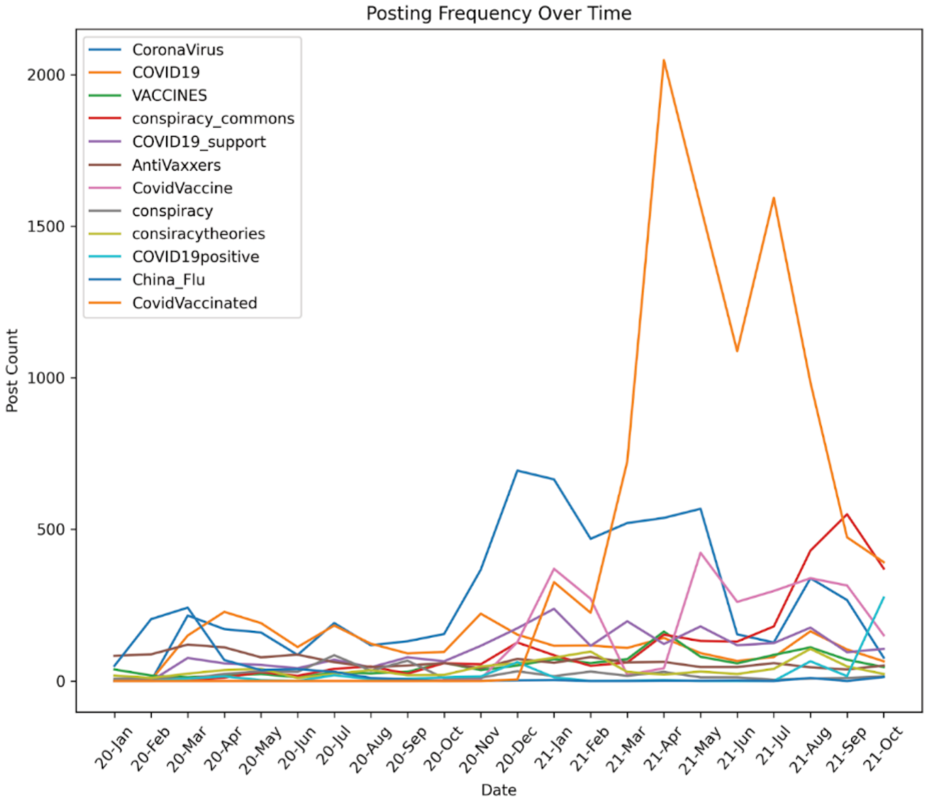


Figure 2: Reddit 1.2 Posting-Frequency Over Time. Subreddits are symbolized by lines of varying colors.

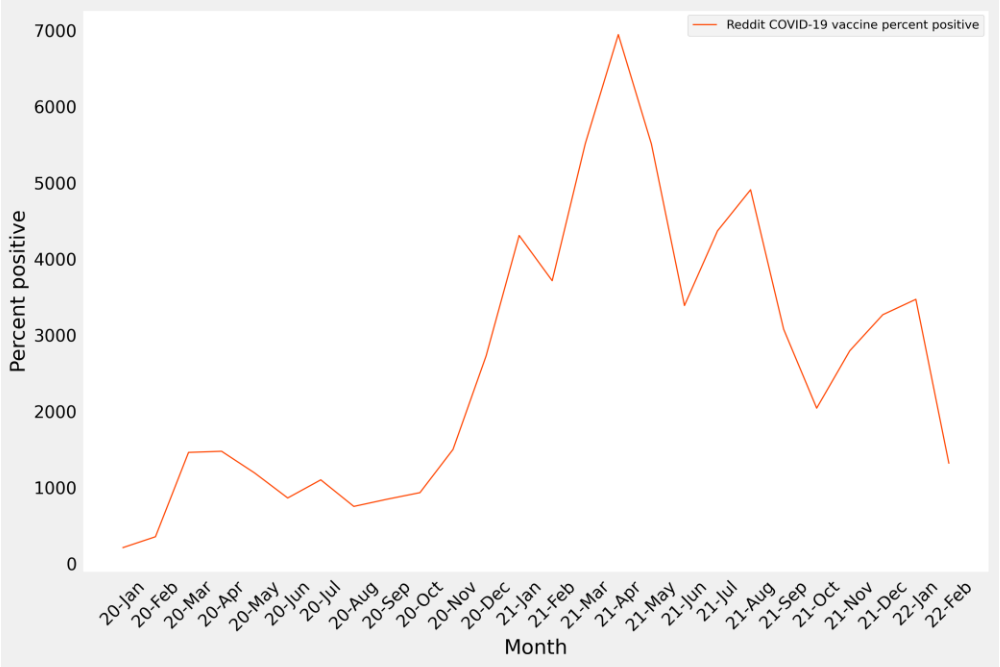


Figure 3: Reddit 2.0 posting frequency from January 1, 2020 to March 1, 2022.

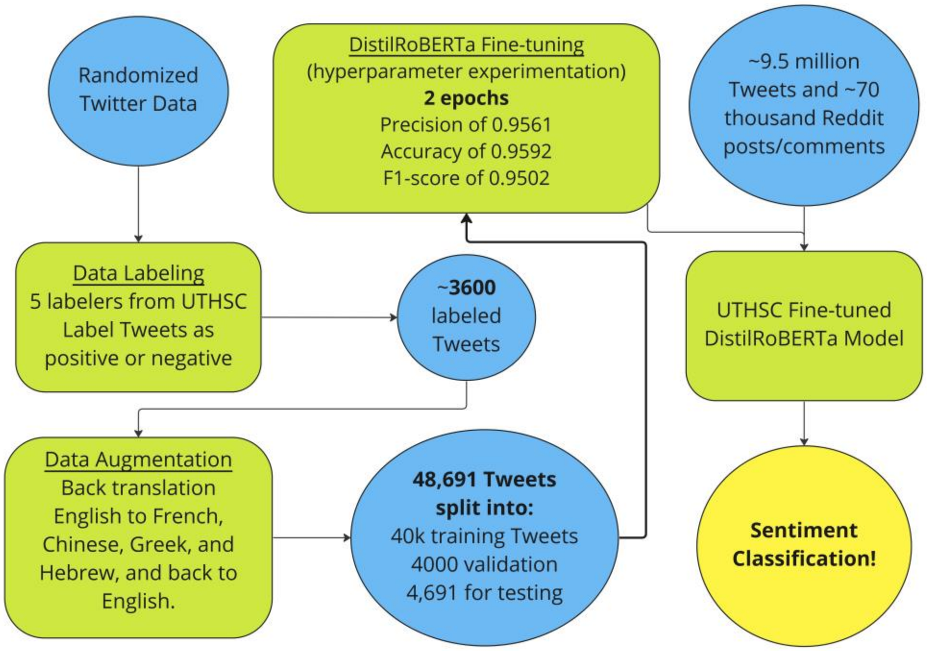

Randomized Twitter Data
DistilRoBERTa Fine-tuning
(hyperparameter experimentation)
2 epochs
Precision of 0.9561
Accuracy of 0.9592
F1-score of 0.9502
~9.5 million Tweets and ~70 thousand Reddit posts/comments
Data Labeling
5 labelers from UTHSC Label Tweets as positive or negative
~3600 labeled Tweets
UTHSC Fine-tuned DistilRoBERTa Model
Data Augmentation
Back translation English to French, Chinese, Greek, and Hebrew, and back to English.
48,691 Tweets split into:
40k training Tweets
4000 validation
4,691 for testing
Sentiment Classification!

Figure 4: Flowchart of DistilRoBERTa sentiment classification.

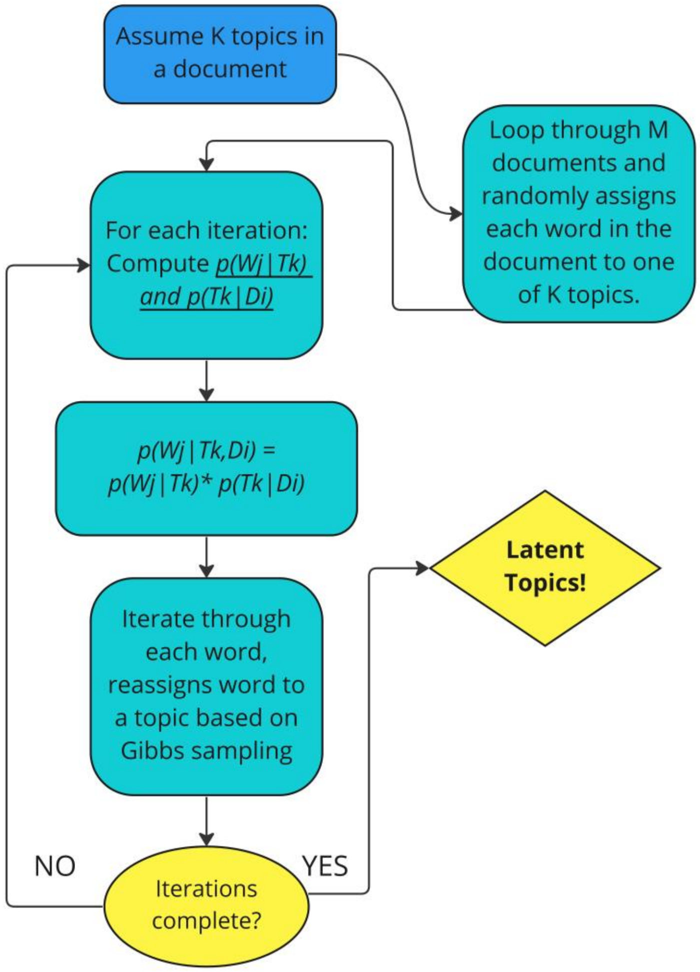

Assume K topics in a document
Loop through M documents and randomly assigns each word in the document to one of K topics.
For each iteration: Compute p(Wj|Tk) and p(Tk|Di)
p(Wj|Tk,Di) = p(Wj|Tk)* p(Tk|Di)
Iterate through each word, reassigns word to a topic based on Gibbs sampling
NO
YES
Iterations complete?
Latent Topics!

Figure 5: Flowchart of the LDA algorithm.

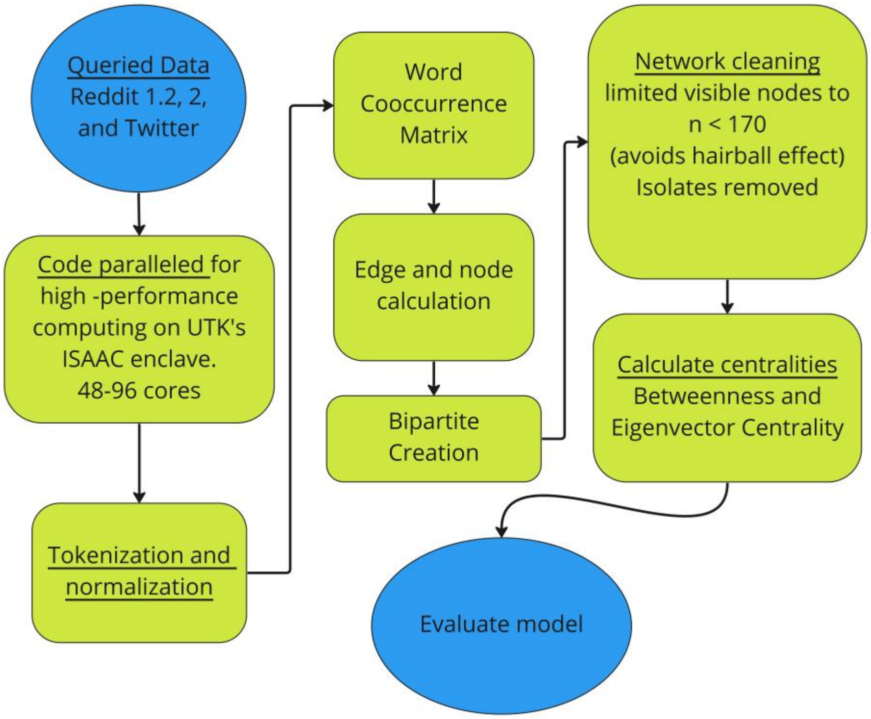

Queried Data
Reddit 1.2, 2, and Twitter
Code paralleled for high -performance computing on UTK's ISAAC enclave. 48-96 cores
Tokenization and normalization
Word Cooccurrence Matrix
Edge and node calculation
Bipartite Creation
Network cleaning
limited visible nodes to n < 170
(avoids hairball effect)
Isolates removed
Calculate centralities
Betweenness and Eigenvector Centrality
Evaluate model

Figure 6: Flowchart of the process to create and evaluate semantic networks.

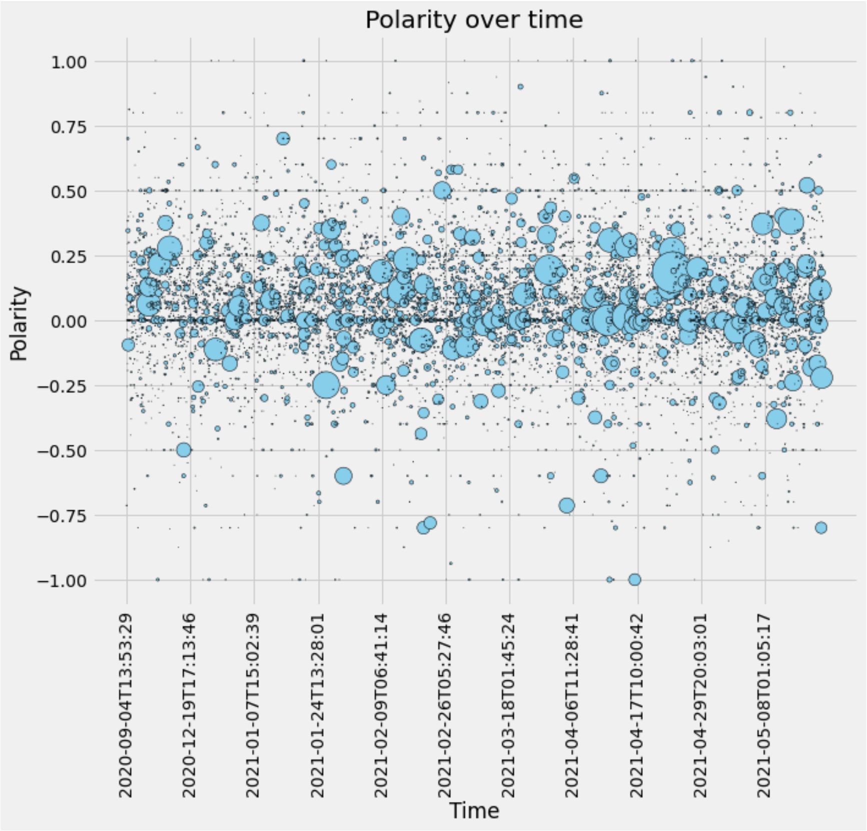


Figure 7: Polarity versus time. Polarity is represented on the y-axis and time is represented on the x-axis. Data points are represented as light blue circles. Circle size indicates the number of upvotes per comment. Data point sizes are reflective of the vote count and are represented by larger circles and smaller quantities are represented by smaller circles.

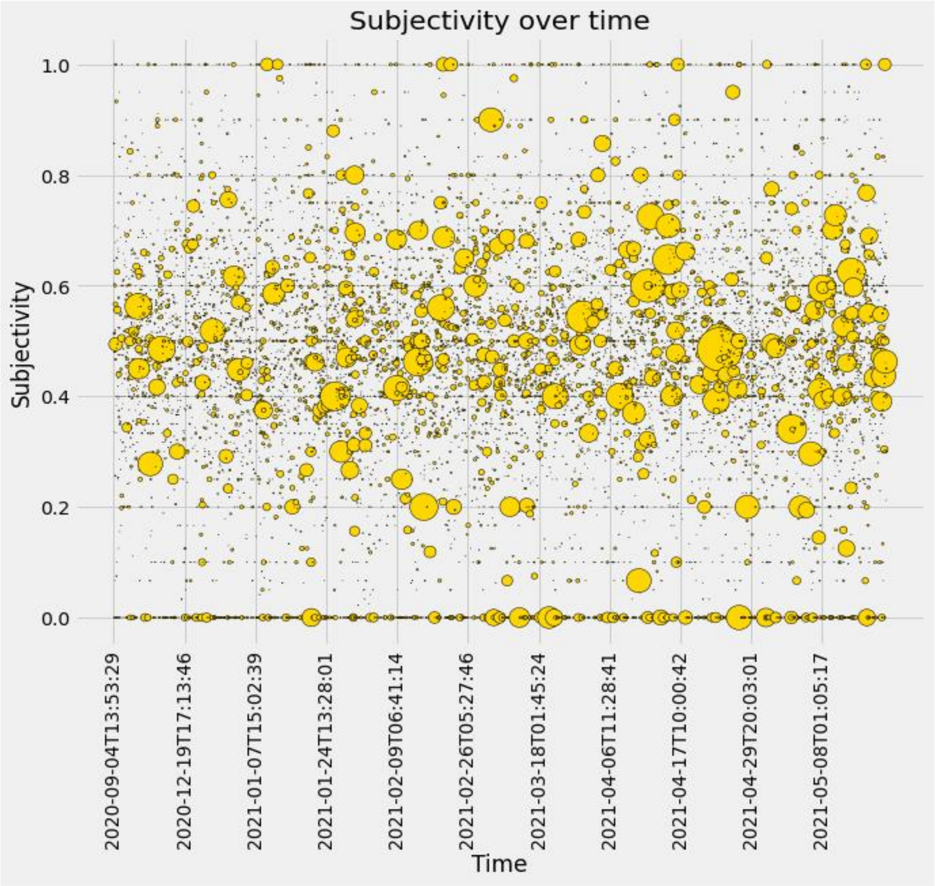


Figure 8: Subjectivity versus time. Polarity is represented on the y-axis and time is represented on the x-axis. Data points are represented as blue circles.

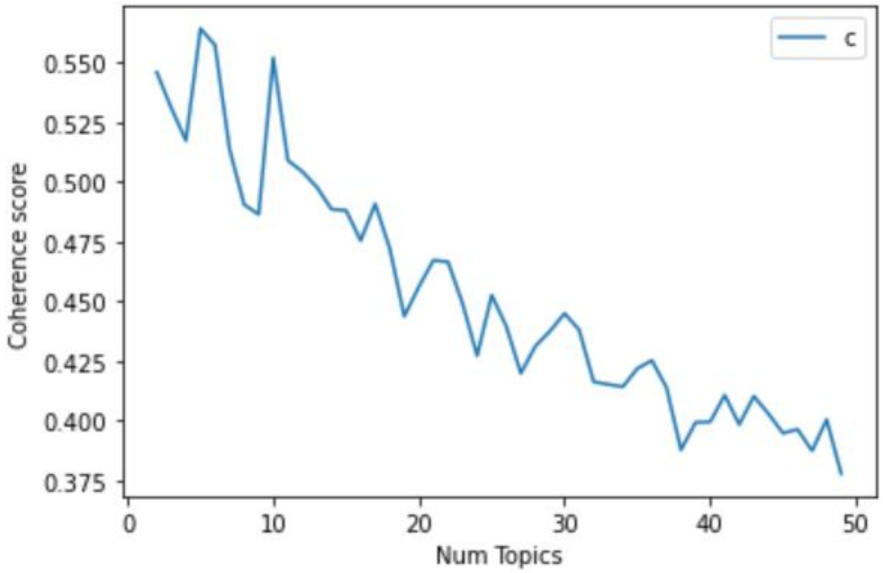


Figure 9: Coherence score vs the number of topics. The y-axis represents coherence and the x-axis represents the number of topics.

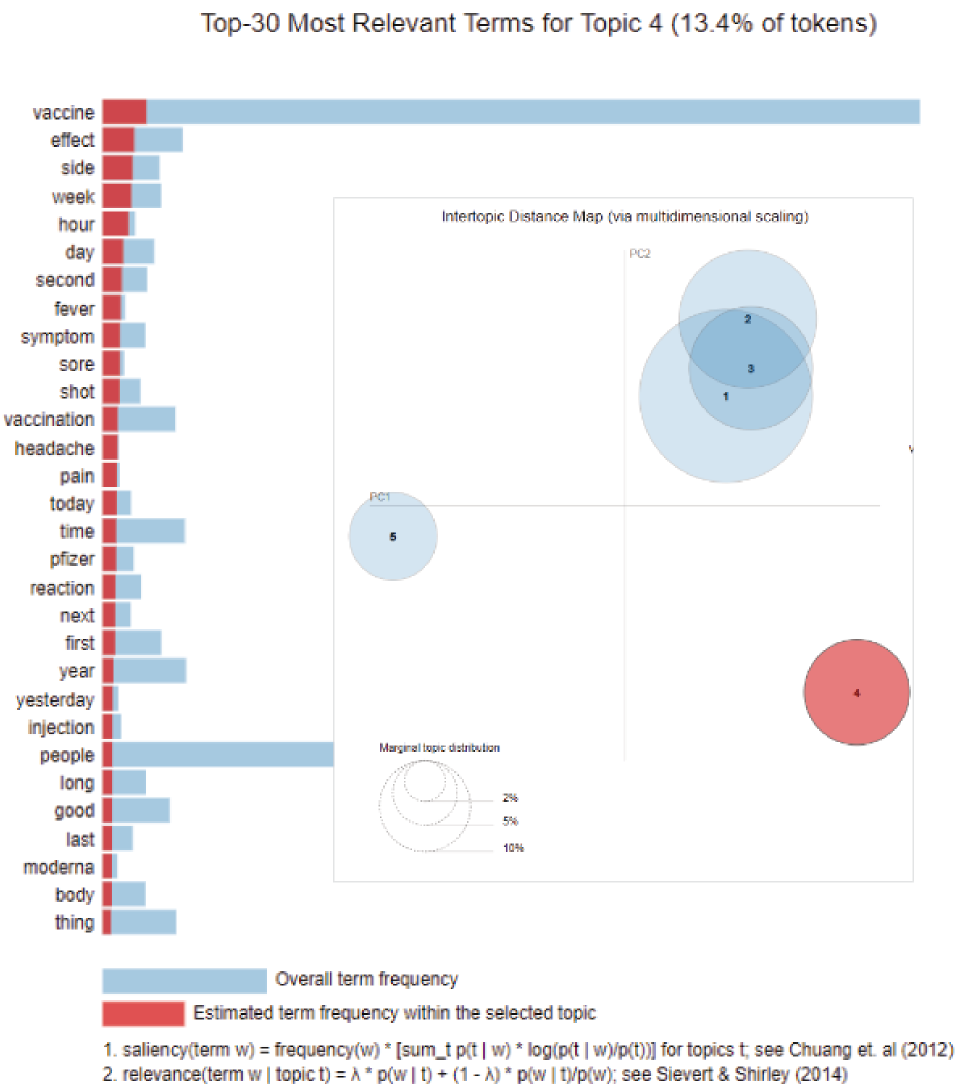


Figure 10: Example of "Topic 4". The blue rectangles are representative of overall term frequency and the red rectangles represent frequency within Topic 1. Please see github.com/Cheltone/NLP_Reddit for an interactive display of LDA topics. Spheres represent relative topic distribution.

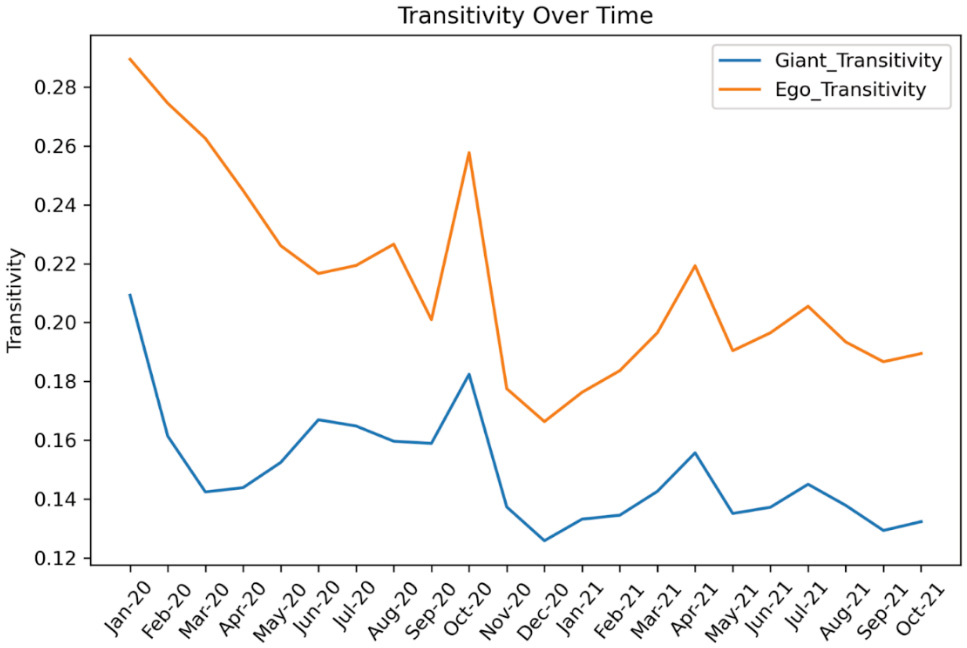


Figure 11: Transitivity Over Time. The orange line represents the *Vaccine* Ego network and the blue line represents the *Giant* network transitivity.

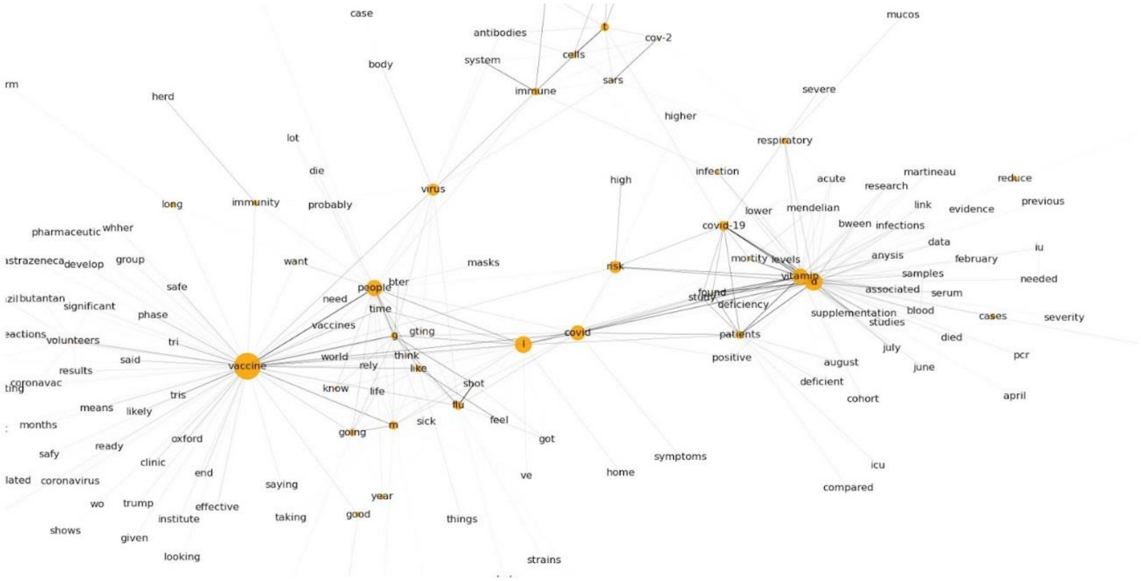


Figure 12: Betweenness Centrality for September 2020. Nodes are indicated by orange circles and edges are indicated by lines. Node size is reflective of the weight and edge thickness indicative of interconnectedness between nodes. Note the *vaccine*-centered cluster and the *vitamin* and *d*-centered clusters are mainly connected by the *covid* node in between the two clusters.

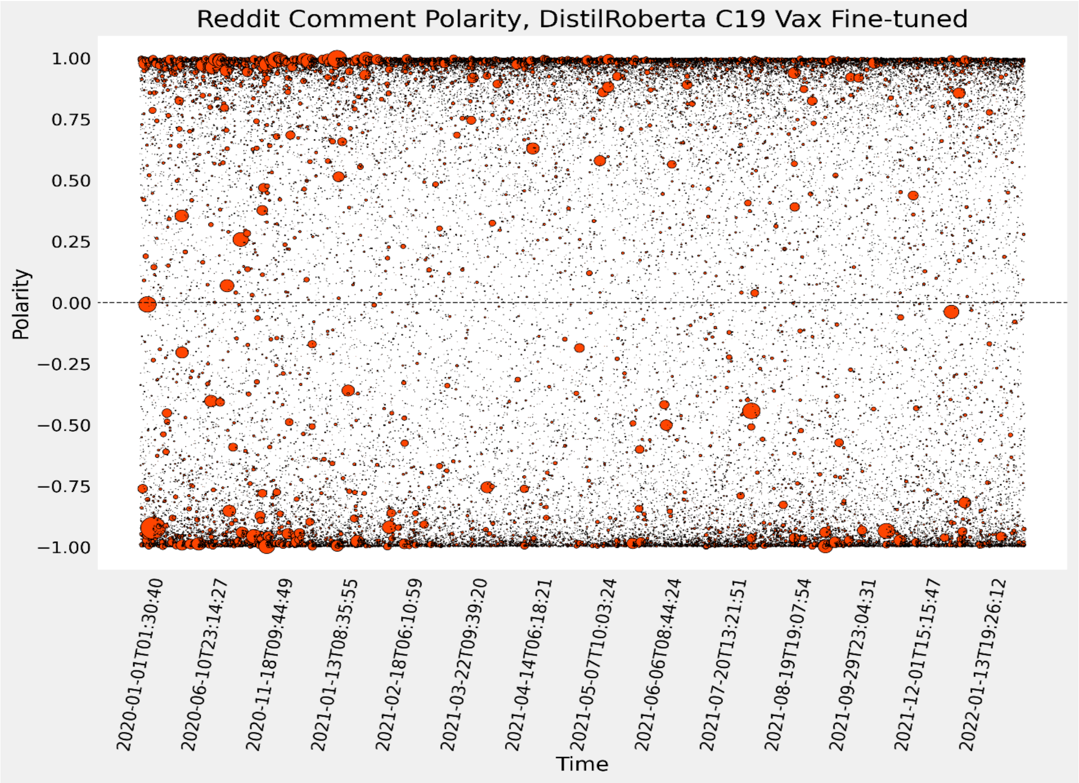


Figure 13: Reddit Comment Polarity for DistilRoBERTa Fine-tuned. Polarity and corresponding confidence probability are represented on the y-axis and time is represented on the x-axis. Data points are represented as orange-red circles. Circle size indicates the number of upvotes per comment. Datapoint sizes are reflective of the vote count and are represented by larger circles and smaller quantities are represented by smaller circles.

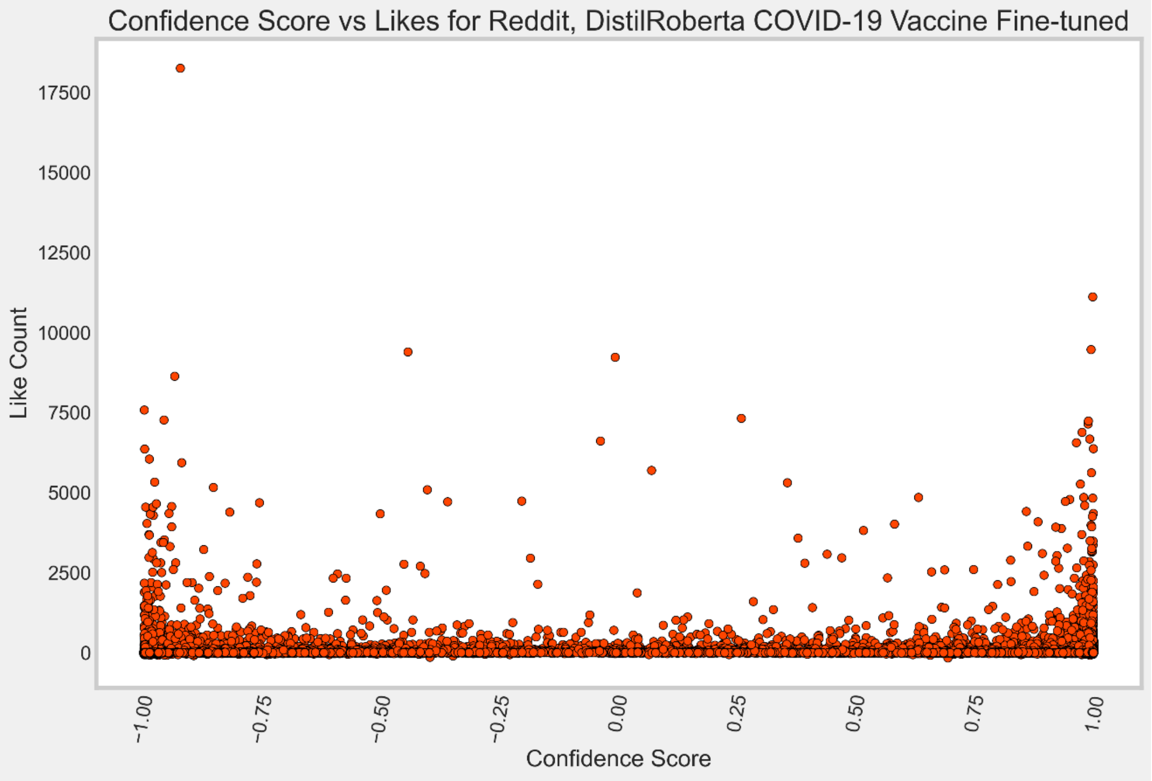


Figure 14: Confidence Score vs Like Count for Twitter. The x-axis represents the confidence score and the y-axis represents the number of likes a Tweet received Data points below 0.0 on the x-axis symbolize a negative classification and points above 0.0 represent a positive classification. Data points are represented as light blue circles.

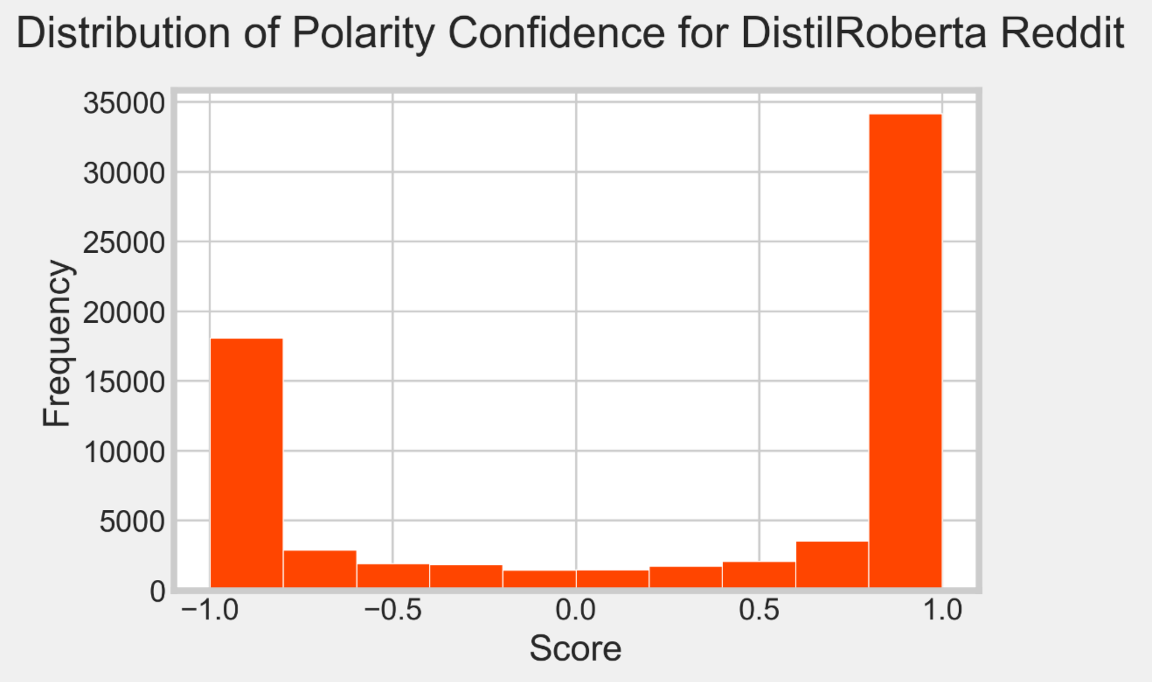


Figure 15: Histogram representing confidence score and number of Tweets. Frequency is equivalent to millions.

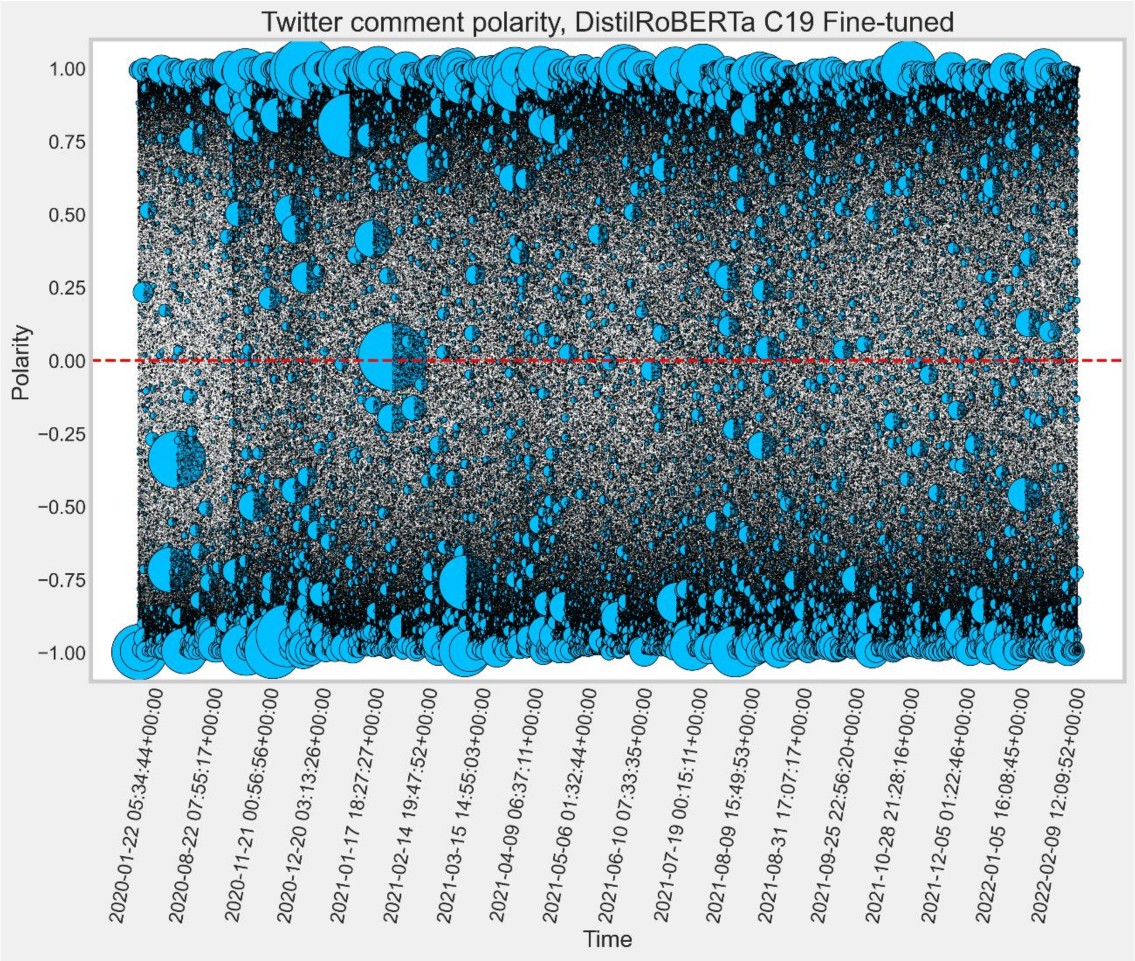


Figure 16: Tweet Polarity for DistilRoBERTa Fine-tuned. Polarity and corresponding confidence probability are represented on the y-axis and time is represented on the x-axis. Data points are represented as light blue circles. Circle size indicates the number of upvotes per comment. Datapoint sizes are reflective of the vote count and are represented by larger circles and smaller quantities are represented by smaller circles.

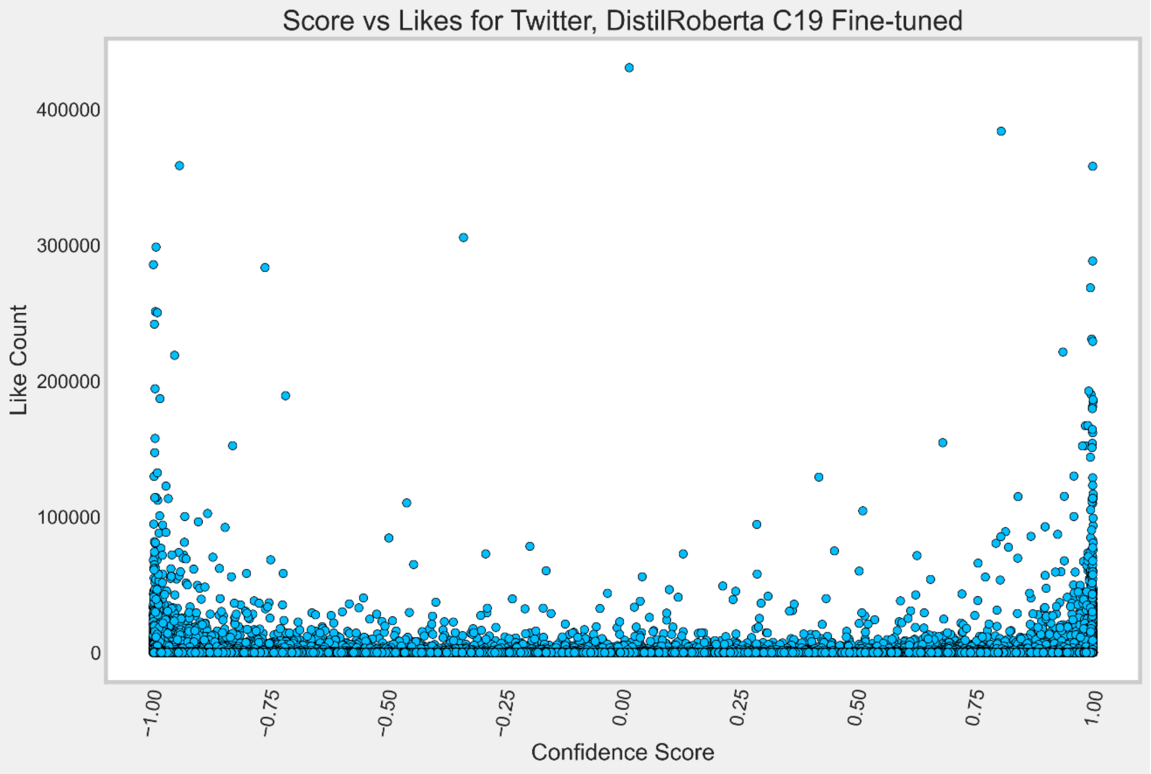


Figure 17: Confidence Score vs Like Count for Twitter. The x-axis represents the confidence score and the y-axis represents the number of likes a Tweet received Data points below 0.0 on the x-axis symbolize a negative classification and points above 0.0 represent a positive classification. Data points are represented as light blue circles.

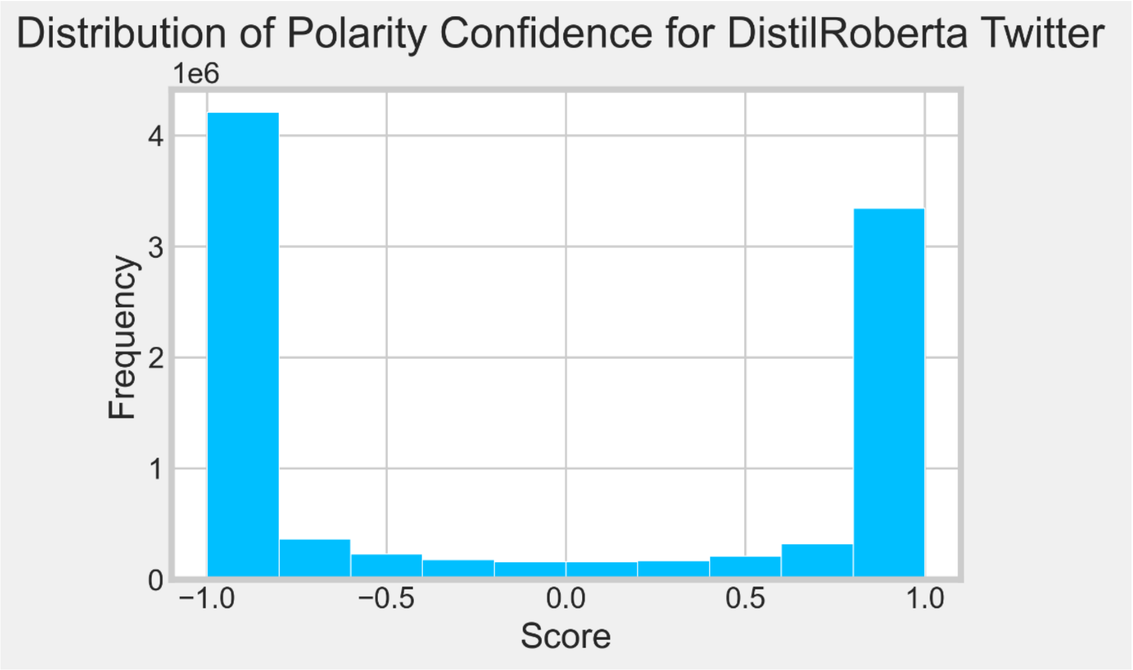


Figure 18: Histogram representing confidence score and number of Tweets. Frequency is equivalent to millions.

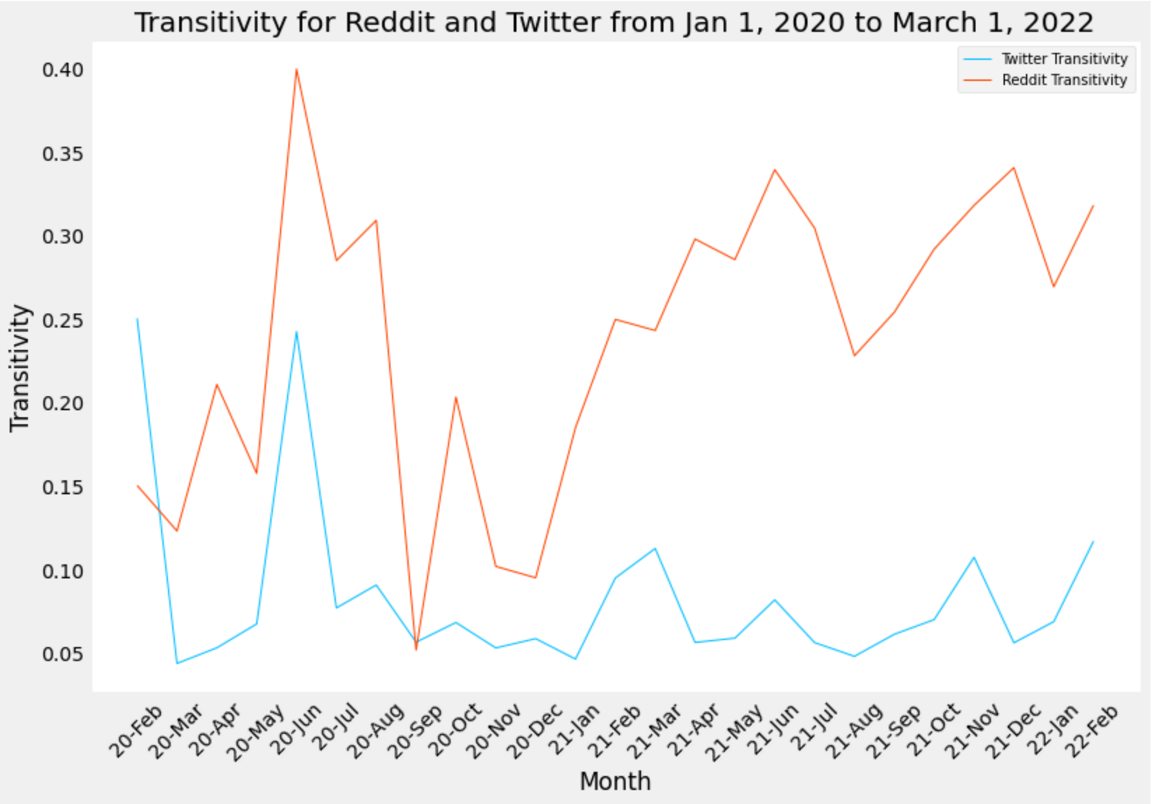


Figure 19: Transitivity Over Time. The orange line represents the *Reddit (Reddit 2.0)* network and the blue line represents the *Twitter* network transitivity across time.

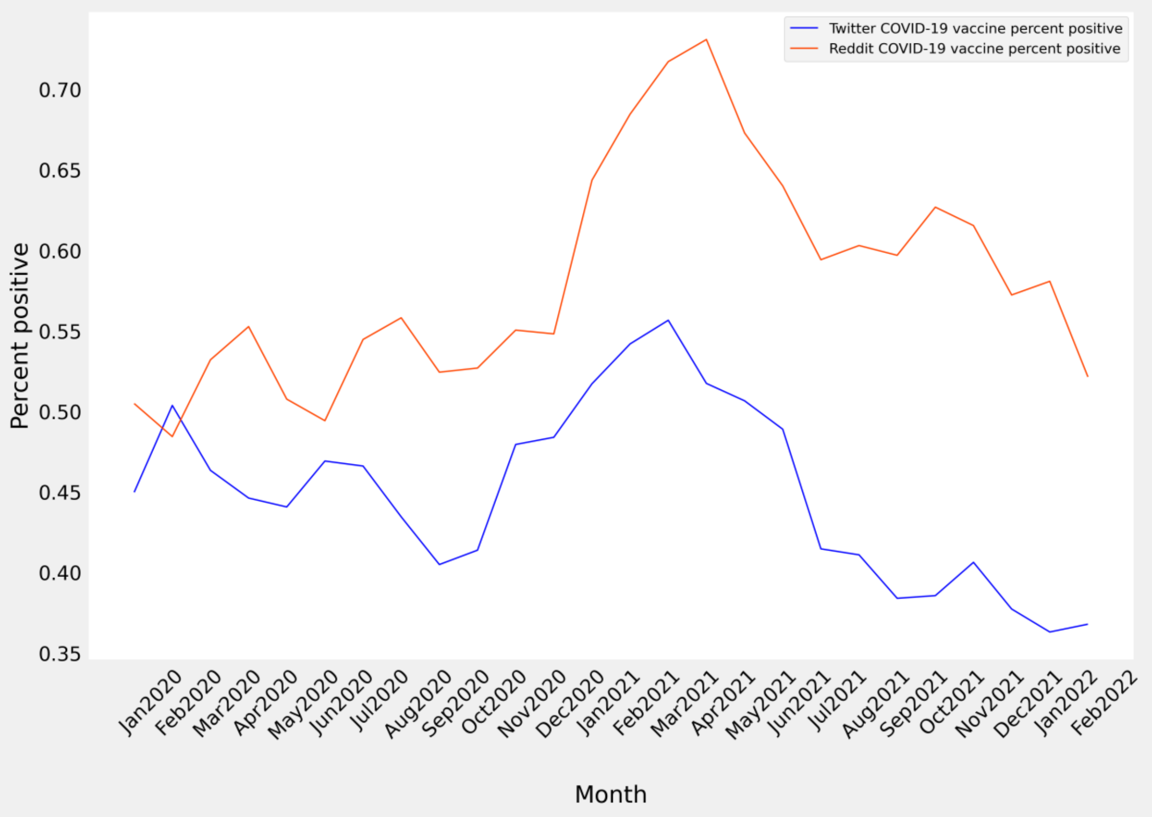
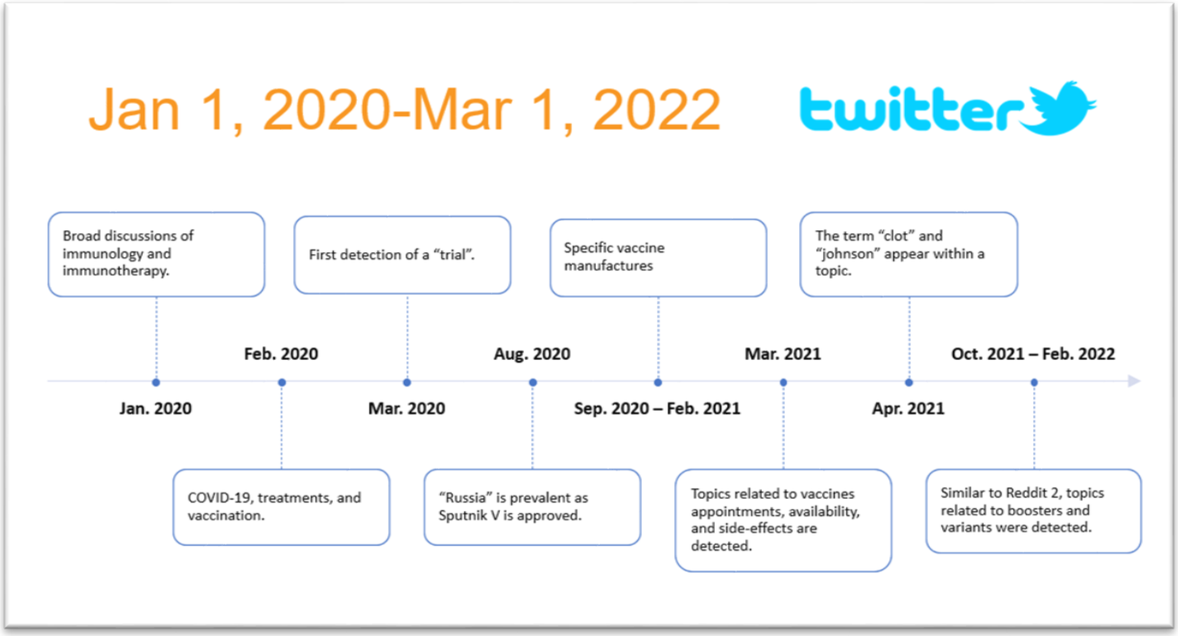


Figure 20: Monthly Sentiment for Twitter and Reddit COVID-19 vaccine-related posts. The x-axis represents time and the y-axis represents the percentage of posts classified as positive. The blue line represents Twitter sentiment and the orange-red line represents Reddit sentiment. Note: Because posting frequency was very low, sentiment for January 2020 is an average of all other months for corresponding data.



Jan 1, 2020-Mar 1, 2022

twitter

Jan. 2020: Broad discussions of immunology and immunotherapy.

Feb. 2020: COVID-19, treatments, and vaccination.

Mar. 2020: First detection of a "trial".

Aug. 2020: "Russia" is prevalent as Sputnik V is approved.

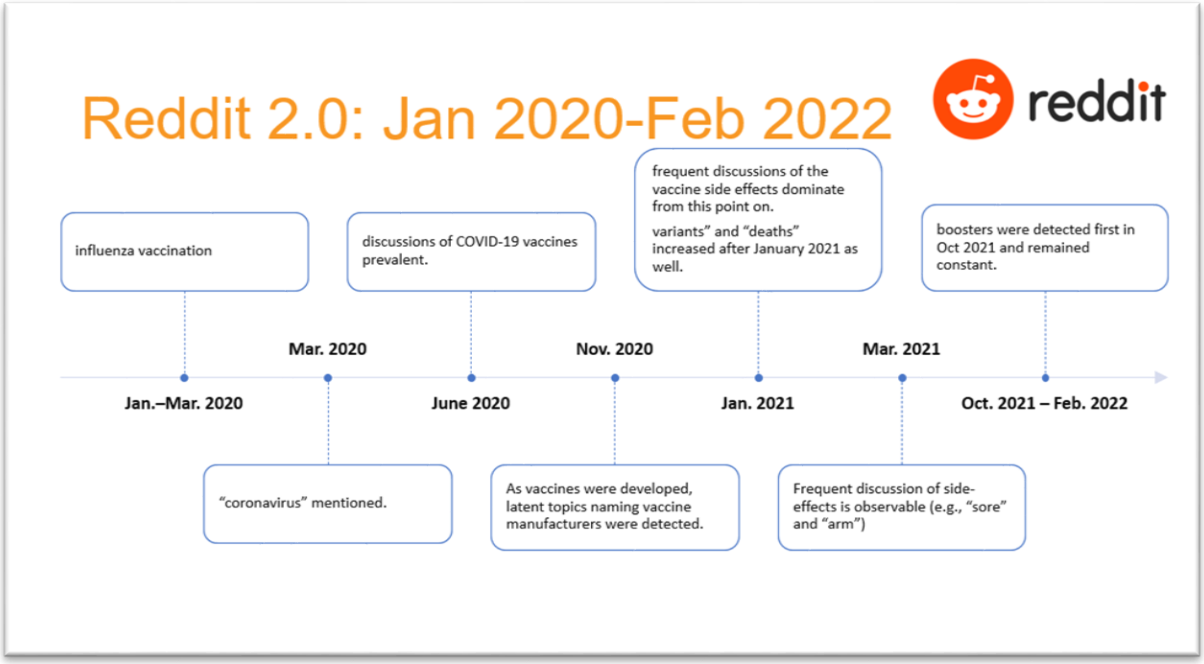
Sep. 2020 – Feb. 2021: Specific vaccine manufactures

Mar. 2021: Topics related to vaccines appointments, availability, and side-effects are detected.

Apr. 2021: The term "clot" and "johnson" appear within a topic.

Oct. 2021 – Feb. 2022: Similar to Reddit 2, topics related to boosters and variants were detected.

Figure 21: Twitter LDA timeline



Reddit 2.0: Jan 2020-Feb 2022

Jan.–Mar. 2020: influenza vaccination

Mar. 2020: “coronavirus” mentioned.

June 2020: discussions of COVID-19 vaccines prevalent.

Nov. 2020: As vaccines were developed, latent topics naming vaccine manufacturers were detected.

Jan. 2021: frequent discussions of the vaccine side effects dominate from this point on.

variants” and “deaths” increased after January 2021 as well.

Mar. 2021: Frequent discussion of side-effects is observable (e.g., “sore” and “arm”)

Oct. 2021 – Feb. 2022: boosters were detected first in Oct 2021 and remained constant.

Figure 22: Reddit LDA timeline

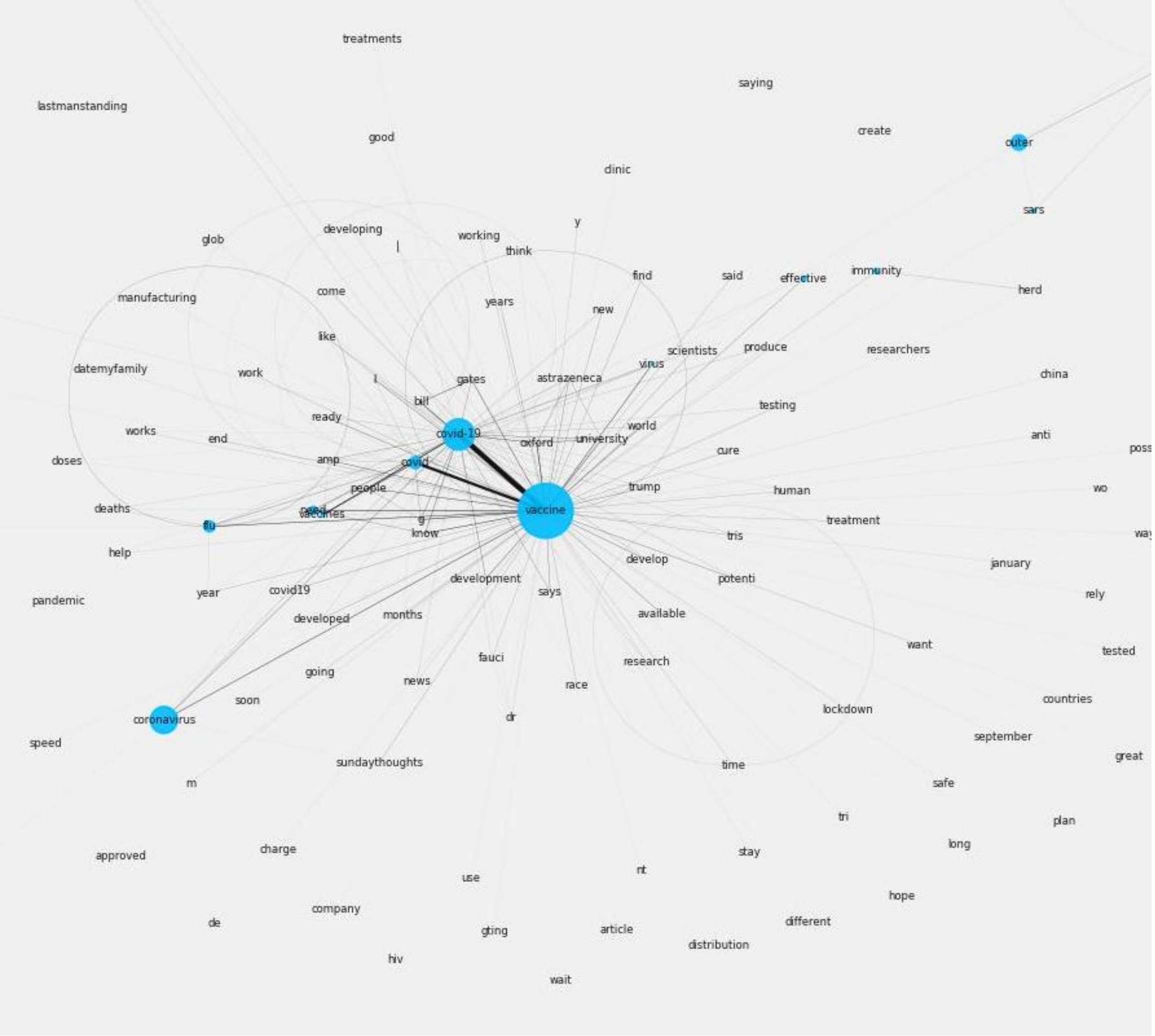


Figure 23: Twitter network April 2020

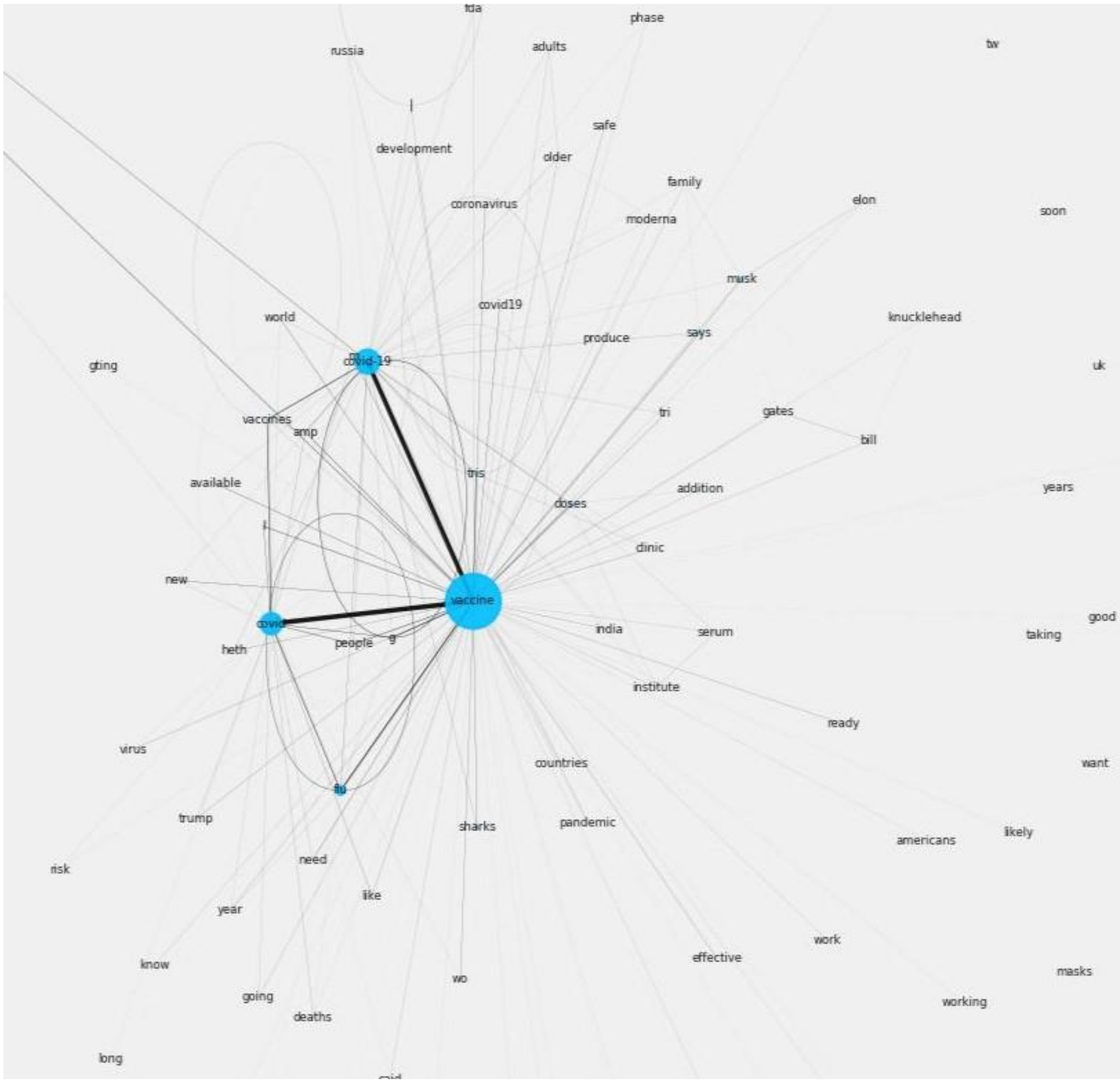


Figure 24: Twitter network September 2020

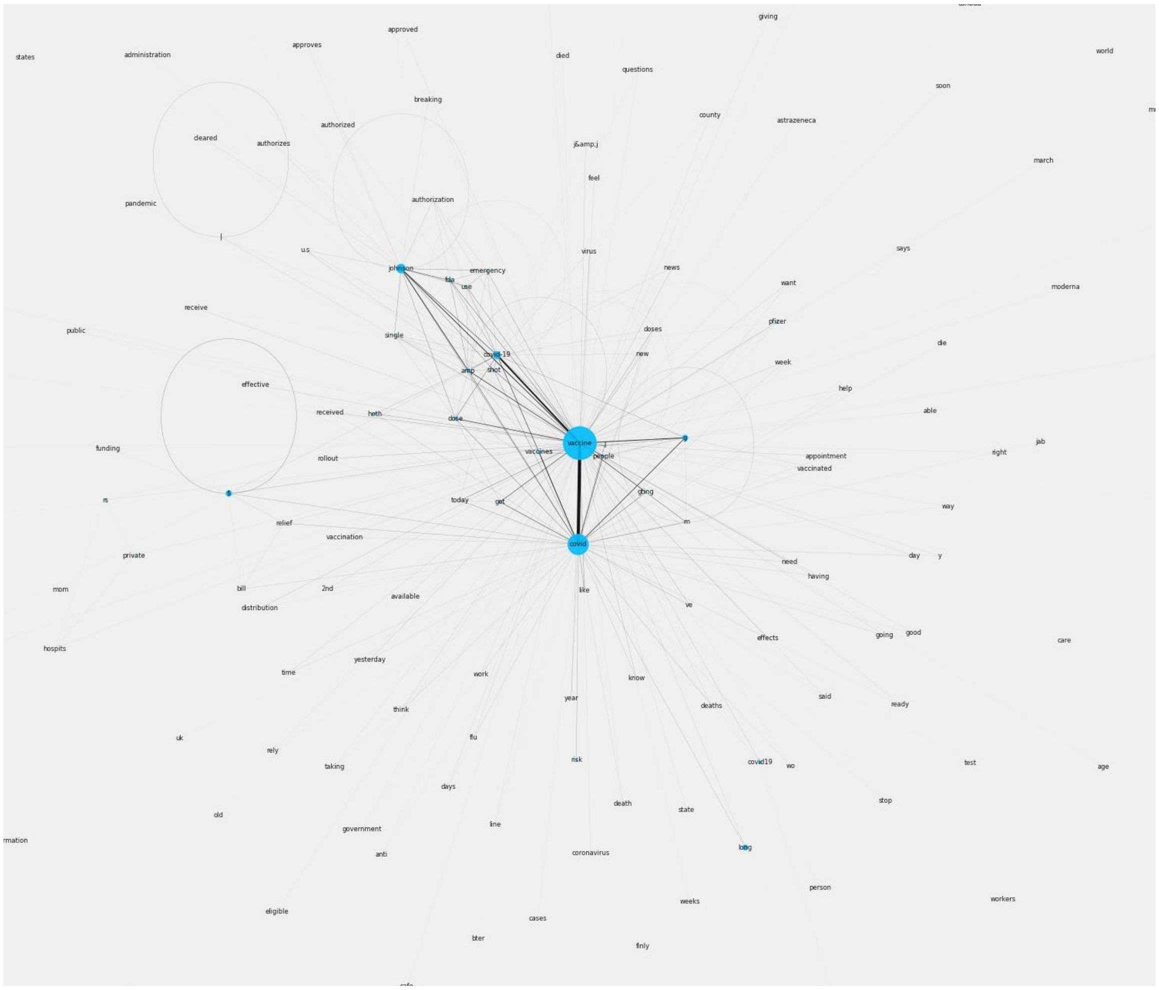


Figure 25: Twitter network February Feb 2021

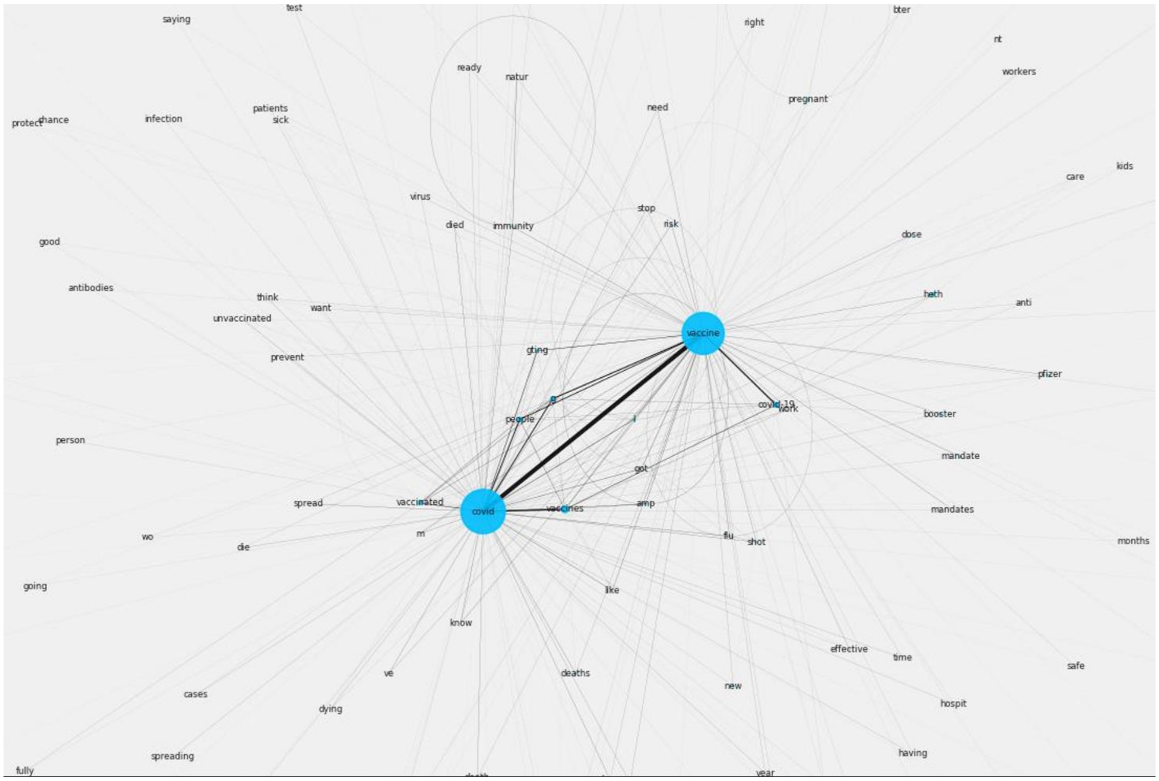


Figure 26: Twitter network September 2021

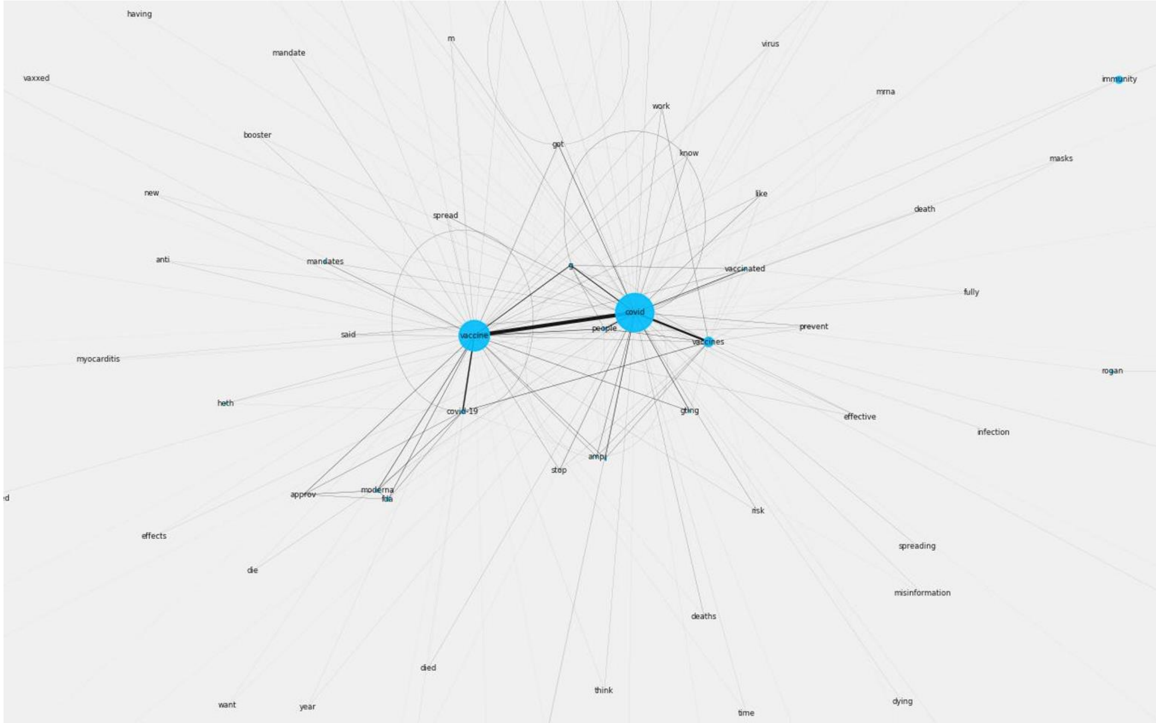


Figure 27:Twitter network January 2022

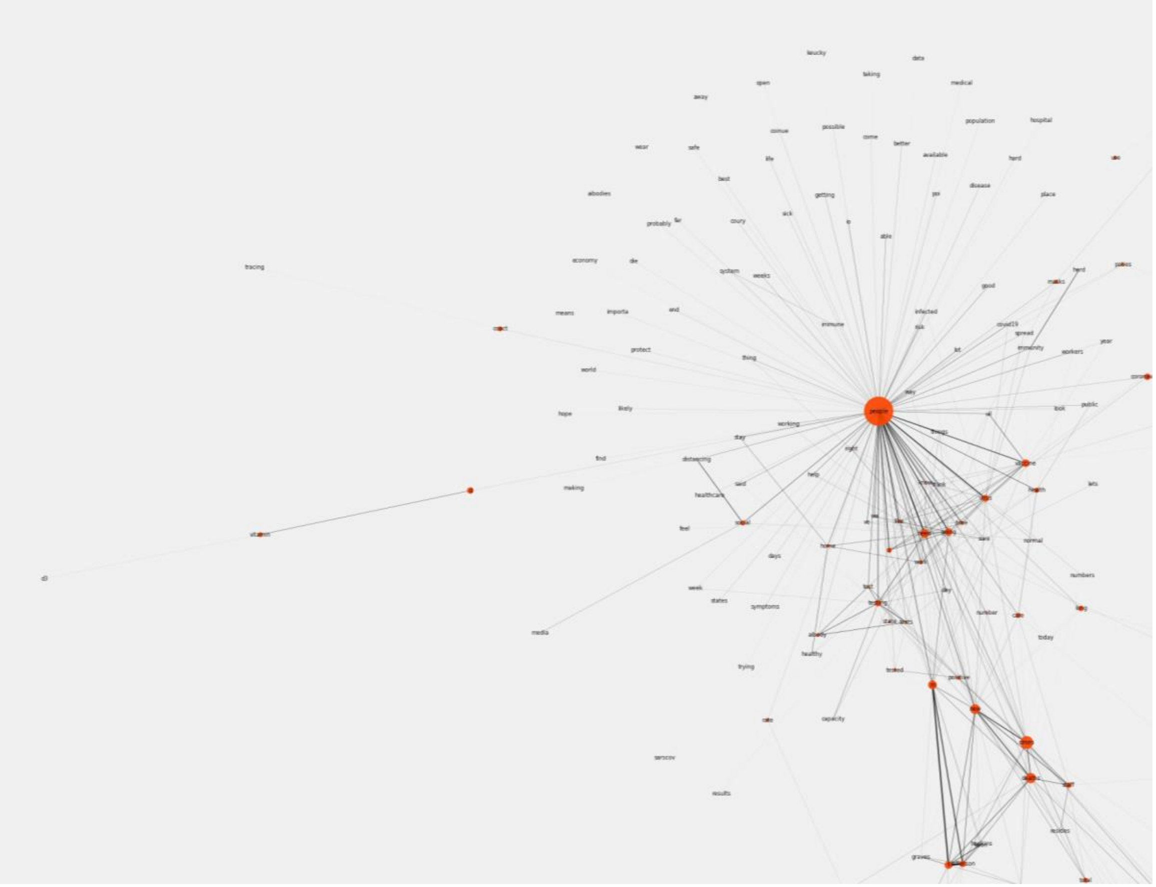


Figure 28: Reddit network April 2020

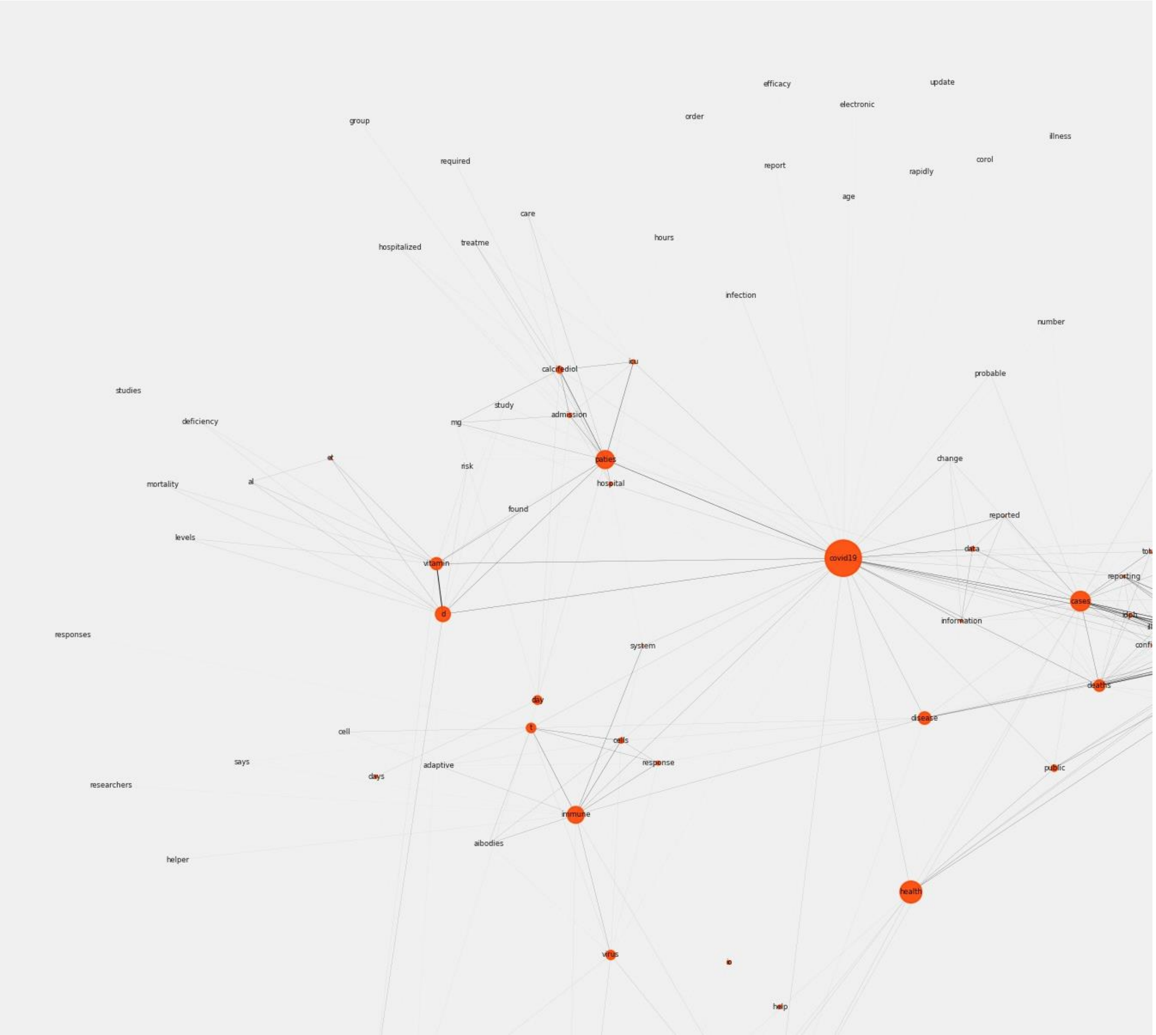


Figure 29: Reddit 2.0 betweenness centrality

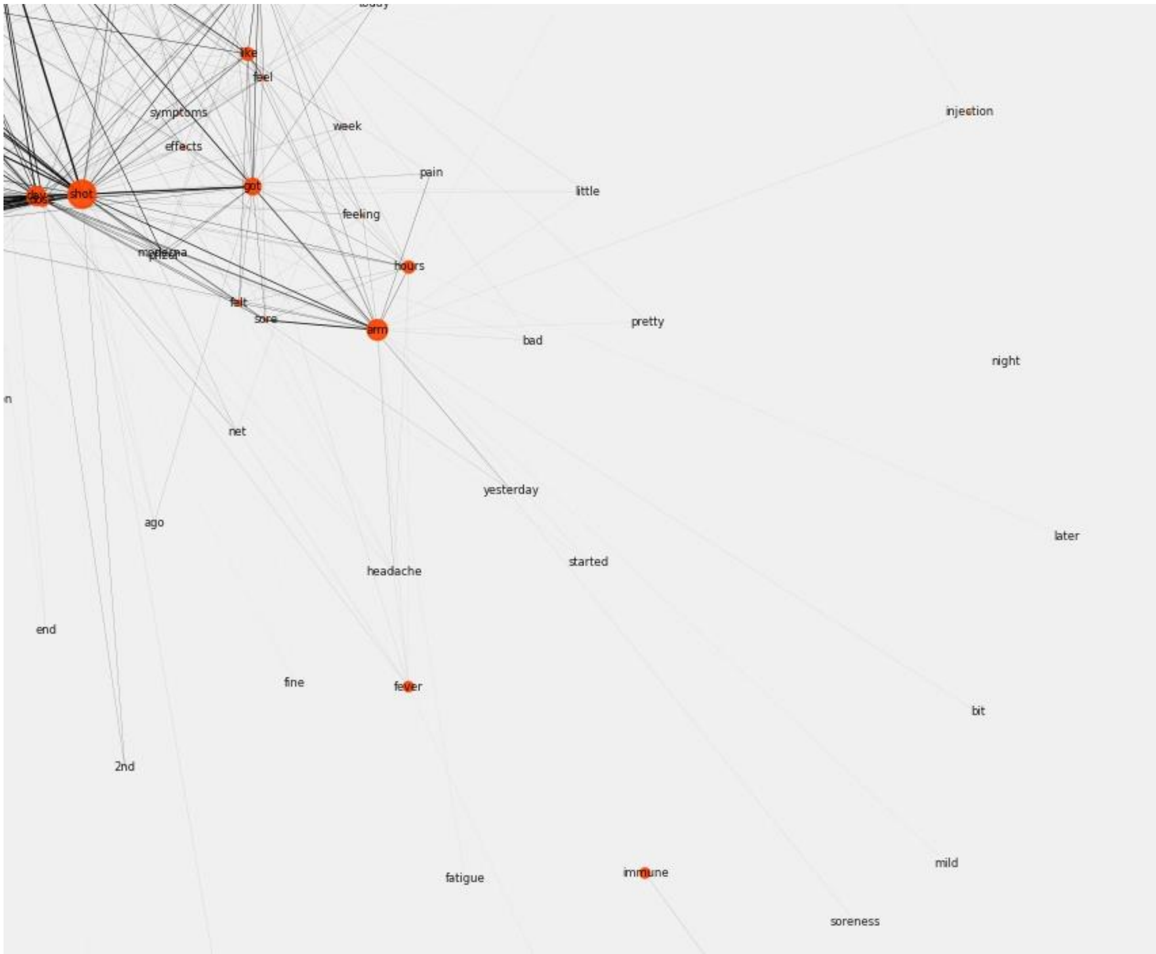


Figure 30: Reddit network March 2021

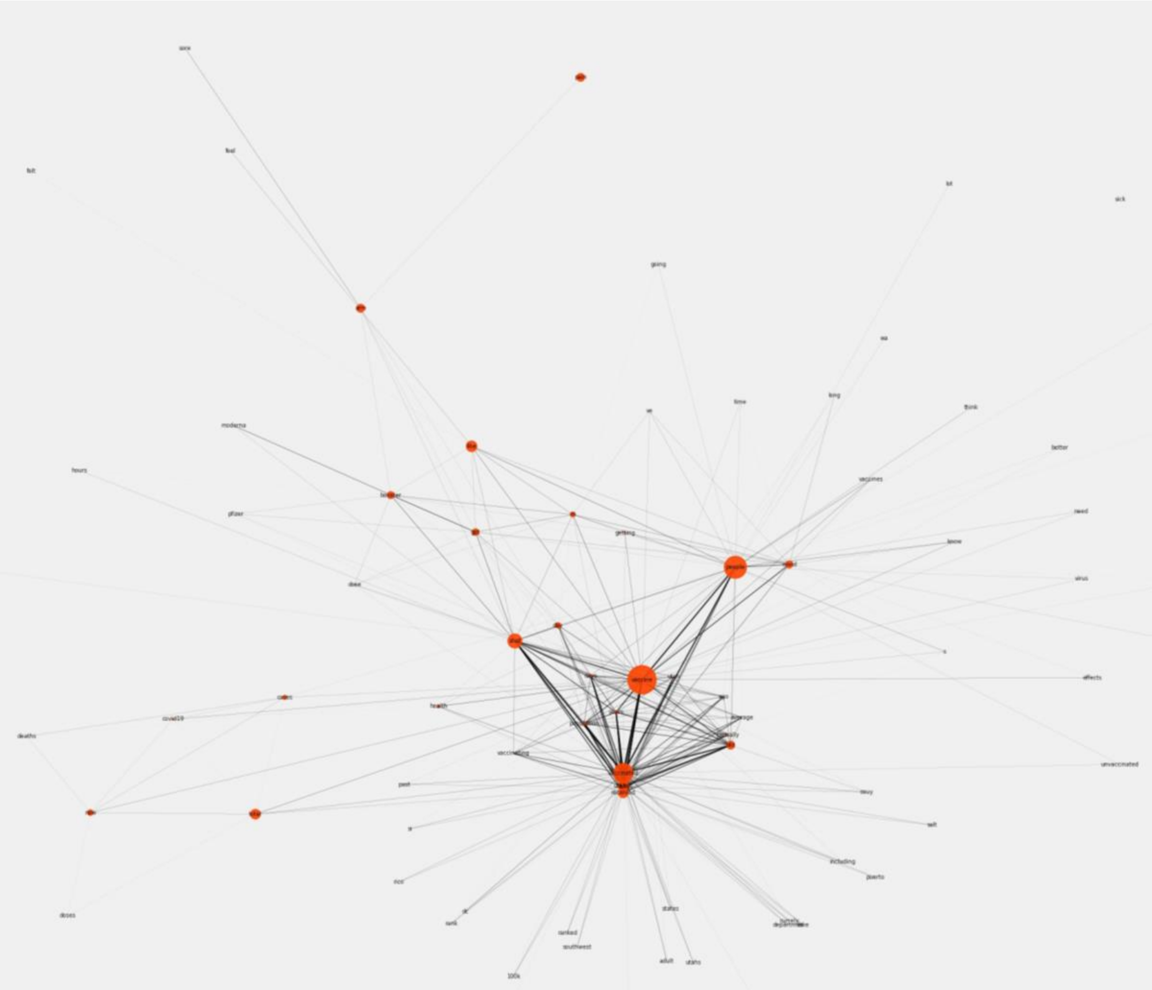


Figure 31: Reddit network November 2021

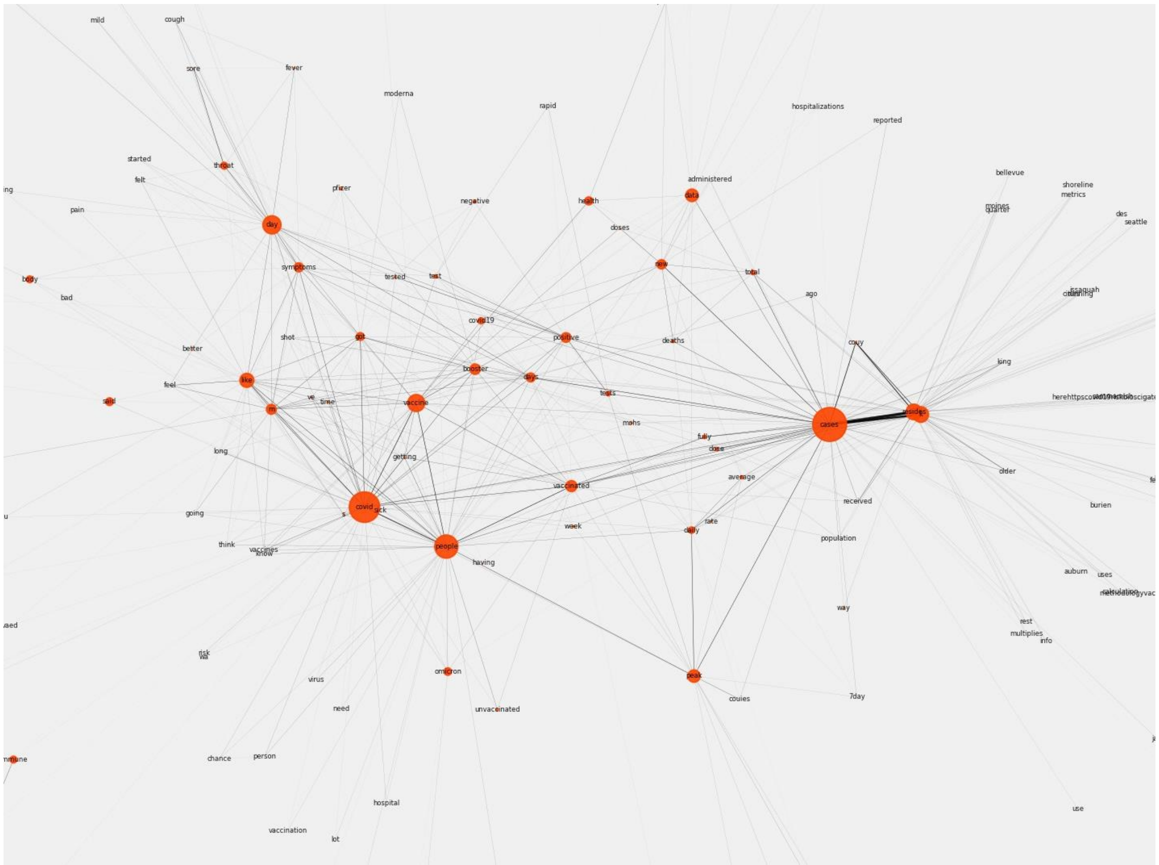


Figure 32: Reddit network January 2022

VITA

Chad Melton was born in Dallas, TX and grew up in East Tennessee. He is currently a Doctoral Candidate in Data Science and Engineering and Bredesen Center Fellow at The University of Tennessee-Oak Ridge Innovation Institute. Chad will continue to explore scientific methods to contribute to the betterment of mankind.