Public Perception of the CDC and the COVID-19 Pandemic through Topic Modeling and Qualitative Analysis

Harrison Brown*

2021-12-02

Abstract

Social media is a powerful source of data regarding individual perception of public health policy and related phenomena such as the COVID-19 pandemic. This study performs topic modeling – a statistical method capable of determining abstract topics within collections of documents – on approximately 300,000 Tweets from January 1, 2019, to September 21, 2021, to better understand user sentiment of potentially conversational topics within the discourse. Topic modeling is performed on Tweets by the CDC's official Twitter account *"@CDCgov"* (*CDC Tweets*), as well as on Tweets that reply to, quote, or mention the CDC's Twitter handle (*Reply Tweets*). Topic modeling is done through use of the *Latent Dirichlet Allocation (LDA)* algorithm, with 11 distinct topics. The results of multinomial logistic regression show the relationship between topics of CDC Tweets and topics of Reply Tweets, leading to a quantification of the public health messages which are deemed divisive or controversial. The regression analysis indicates that the most "controversial" Tweets authored by the CDC were those relating to mask wearing and mask mandates, with significantly higher occurrences of replies regarding "Big Pharma", government skepticism, and vaccine efficacy skepticism. The results of this study indicate a widespread mistrust in governmental organizations, public health policies, and science communication in general; however, the methodology used can be applied to public health messaging, allowing organizations to better communicate with the public.

*Appalachian State University, brownhr@appstate.edu

Contents

1 Introduction

Within the United States, recent trends have shown a large-scale decrease in public trust of not only government organizations (e.g., the CDC and FDA), but also science as a whole (Muñoz, Moreno, and Luján 2012; Huang and Liu 2020; Choli and Kuss 2021). This skepticism and widespread mistrust are dangerous when placed into the context of ignorance regarding COVID-19 and public safety measures. Using social media (specifically, Twitter) as a source for public opinion, this paper aims to study the effects of messaging and public outreach measures by the CDC on public opinion and perception, with the intention of improving public health communication.

1.1 Research Question

This study intends to answer the following research question; how can qualitative text mining methods be used to better understand public opinion, allowing for more effective public health messaging and science communication. In answering this question, this paper aims to develop the foundation of a method for predicting discussion topics on social media using *Latent Dirichlet Allocation (LDA)* and regression analysis to increasing effectiveness of science communication programs. Much of the literature focuses on the societal effects and public perception of the COVID-19 pandemic *itself* (Boon-Itt and Skunkan 2020; Dubey 2020; Manguri, Ramadhan, and Amin 2020; Zulfikar and Auliansyah 2020; Garcia and Berton 2021). This study instead aims to measure user opinion from the perspective of public health geography by examining the relationship between the topics of Tweets by the CDC (defined as *CDC Topic* and *CDC Tweet*) and their effects on the topics of Tweets made in reply (*Reply Topic* and *Reply Tweet*).

2 Literature Review

While much research regarding text mining of Twitter data has focused specifically on the COVID-19 pandemic, few works have focused on examining the public perception of government entities within the social media discourse (Dubey 2020; Boon-Itt and Skunkan 2020; Manguri, Ramadhan, and Amin 2020; Garcia and Berton 2021). By focusing instead on governmental organizations, it is possible to reach an understanding of why public health messaging has not achieved its intended goals within the U.S., such as with vaccine hesitancy, COVID-19 denial, and a shifting of the discourse towards "personal freedom". This study aims to provide both a qualitative and quantitative analysis of public perception, using the literature below as a foundation.

2.1 Sentiment Analysis

Much research regarding public opinion on the COVID-19 pandemic uses *sentiment analysis*, a dictionary-based approach to quantifying user sentiment (i.e. emotional valence); while this approach is useful for exploratory analysis of user sentiment, shortcomings exist when applying this model to "short-text" formats such as social media microblogging, comments, and Twitter posts (Pak and Paroubek 2010; Liu 2012; Clavel and Callejas 2015; Boon-Itt and Skunkan 2020; Dubey 2020; Manguri, Ramadhan, and Amin 2020; Garcia and Berton 2021). Due to these shortcomings (i.e. because many Tweets in this corpus contain fewer than 10 words), sentiment analysis was not performed within this study, with the focus instead placed on abstract connections between topics generated by topic modeling described below in Section 4.1.3.

Because sentiment analysis obtains scores of emotional valence by referencing a *sentiment dictionary*, context is often lacking in such analyses; phrases such as "good" and "not bad" are semantically similar when understood through natural language, but would receive possible scores of $+1$ and -2,

respectively. Methodology exists for alleviating such issues, including Pak and Paroubek (2010), the results of sentiment analysis are not applicable or necessary for this study.

2.2 Topic Modeling

Extensive literature exists for topic modeling of Twitter data – most recent research studies public perception of the COVID-19 pandemic and its societal effects (Bao et al. 2009; Hong and Davison 2010; Alghamdi and Alfalqi 2015; Debortoli et al. 2016; Negara, Triadi, and Andryani 2019; Boon-Itt and Skunkan 2020; Garcia and Berton 2021). While topic modeling is a powerful text mining tool, some limitations exist when used on *short-text* data; many metrics used to quantify the effectiveness of *LDA* models, such as *coherence score*, do not apply as readily to *short-text* formats. For this study, much of the literature involved foundational theory and methodologies for implementing *LDA* within text mining (e.g., with the `textmineR` R package), as well as methods for data preprocessing when working with *short-text* Twitter data (Alghamdi and Alfalqi 2015; Debortoli et al. 2016; Jones 2019; Rashid, Shah, and Irtaza 2019). Many studies performed topic modeling on Tweets containing keywords related to the COVID-19 pandemic, such as *coronavirus*, *covid*, *pandemic*, etc. (Boon-Itt and Skunkan 2020; e.g., Garcia and Berton 2021).

3 Data

The data for this study consists of a corpus of approximately 300,000 Tweets regarding the Centers for Disease Control and Prevention (CDC). As the CDC disseminates information regarding vaccinations, face coverings (masks), and general public health, tweets made in reply to the CDC (Reply Tweets) often follow the same topic of discussion. The data collection methods defined below explain the workflow for obtaining and cleaning the text corpus within R.

3.1 Twitter API

Twitter data acquisition on a large scale is made possible through the Twitter API v2 Academic Research Access, a platform designed by Twitter for use in academic and scientific research (Twitter API Documentation 2021). Academic access differs from standard API access in that rather than limiting query results to only the last 7 days of public Tweets, the Academic Research Access allows for full-archive searching, as well as much a higher limit on the number of monthly Tweets able to be requested.

3.2 academictwitteR

The `academictwitteR` package for R allows the user to access the Academic Research archive from the Twitter API v2 and was developed in Barrie and Ho (2021). CDC Tweets were gathered by querying "`from:CDCgov`" from Jan. 1, 2019, to Sep. 21, 2021; Reply Tweets were acquired from the same time frame with the query "`@CDCgov`", meaning the text of the Reply Tweet contained the phrase "@CDCgov", the name of the CDC's Twitter handle.

3.3 Conversation ID

In addition to each Tweet's ID, `tweet_id`, Tweets obtained from the `academictwitteR` R package also contain a unique identifier, `conversation_id`, which is an attribute generated by Twitter to keep track of conversation threads, allowing for effective modeling and network analysis on "nonlinear" conversations. (Barrie and Ho 2021; Twitter API Documentation 2021).

3.4 Data Preprocessing

As the corpus contains raw, unfiltered text, steps must be taken to clean and process the data into a format suitable for topic modeling; the methods for processing the text data are outlined below. These steps are performed to ensure accurate results from

topic modeling; for example, the CDC's Twitter handle *@CDCgov" is, by definition, included in all Reply Tweets; this token is removed from the analysis so that the results from topic modeling are not affected.

3.4.1 Tokenization

Tokenization refers to the act of splitting a length of text (a `string`) into individual segments or words, which are most often split by punctuation and whitespace. These tokens consist primarily of unigrams (individual words), but can also be extended to bigrams (two words occurring in order together). Trigrams (three words in-a-row) were omitted as, while they offer significant contextual and semantic information, they occurred too infrequently to be particularly useful; this is a noted issue in *short-text* datasets such as Twitter (Bao et al. 2009; Hong and Davison 2010; Ostrowski 2015; Debortoli et al. 2016).

3.4.2 Stopword Removal

Stopwords are often defined as the most frequently used words which do not offer very much useful information, such as "the", "and", or "at", etc. Stopwords were removed using the `textmineR` and `stopwords` R packages; additionally, hyperlinks and Twitter handle mentions were considered "stopwords" in this analysis as they do not offer semantic value. Other Twitter-specific stopwords include tokens like *RT* ("Retweet"), *amp* (a mis-encoding of the & character), *http*, and *tco*.

Once stopwords, links, and Handle mentions were removed, many Tweets (i.e., spam Tweets) did not contain any text and were thereafter removed. Additionally, extraneous duplicate Tweets by the same author replying to the same CDC Tweet were removed.

3.4.3 Lemmatization

Lemmatization, a form of word stemming, is the process of "reducing" a declined or conjugated form of a word to its base form (e.g., *running*, *runs*, and *run* all reduce to the base form *run*). This is done to preserve the semantic meaning of words, while maintaining a consistent "dictionary" of tokens; for example, with lemmatization, *mask mandates* and *mask mandate* produce the same lemmatized bigram, `mask_mandat`, which allows for comparison of terms of the same effective meaning which differ only in tense or grammatical number. Lemmatization was performed with use of the `SnowballC` R package, which provides several word-stemming algorithms; the default, `porter` is used in this study (Bouchet-Valat 2020).

4 Methodology

The primary methodology for this study involves the use of Topic Modeling, outlined below, to obtain discussion topics for CDC Tweets and Reply Tweets. The distributions and occurrences of these topics can be modeled using *multinomial logistic regression*; the assumption is made that there is a causal relationship between the Topic of a CDC Tweet and the Topic of a Tweet made in direct reply, as the user must understand the semantic meaning behind a given Tweet in order to make a reply.

4.1 Topic Modeling

Topic modeling is a powerful tool set within the field of text mining that allows the user to extract a set of "topics" which occur within a set of documents (i.e. a *corpus*). These topics are based primarily on word co-occurrence; that is, words that appear frequently together are more likely to be assigned to the same topic. For example, because words such as "mask" and "mandate" frequently co-occur as bigrams in discussions on health and sanitation, they are likely to be assigned to the same topic; see table 1 for the top 5 terms within each topic. For this analysis, the topic modeling method used is *Latent Dirichlet Allocation*, which allows for documents to be categorized into more than one topic; see Section 4.1.3 for more detail.

4.1.1 Constructing Corpus

A corpus is defined as a collection of documents for use in text mining; in this case it is the collection of Tweets obtained from the `academictwitteR` package, which were cleaned using the methods described in Section 3.4. This corpus was stored within R as a `data.frame` object, which contained information such as text, the date at which the Tweet was written (`created_at`), the unique ID (`tweet_id`), conversation id (`conversation_id`), and many others.

4.1.2 Document-Term Matrix

A *Document-Term Matrix* (*DTM*) is a construct that represents the occurrences of tokens within each document; as standard, columns in a *DTM* represent each document, rows represent each token, and the values within each cell show the frequency of a given token within a given document. This construction is useful as it allows for a representation of which tokens appear frequently across the entire corpus, and which tokens occur only in a small subset of documents. One major limitation of *DTM*s is the issue of matrix sparsity; the size of the matrix grows exponentially as new terms and documents are added to the corpus, but most cells within the *DTM* have a frequency of 0; the *DTM* generated in this study reported a sparsity of 100%, with rounding errors.

4.1.3 Latent Dirichlet Allocation

Latent Dirichlet Allocation (hereafter *LDA*), developed in Blei, Ng, and Jordan (2003), is an unsupervised approach to topic modeling, in which topics are assigned through "fuzzy clustering" into different subsets of topics. *LDA* allows for *unsupervised* topic modeling – although the number of topics can be defined, the topics themselves (e.g., "mask wearing", "politics") are not known beforehand. Rather than in "hard clustering" algorithms such as hierarchical or k-means clustering, where documents consist of only a single topic, *LDA* assigns a distribution of topics to each document. For example, a Tweet discussing effectiveness of the COVID-19 vaccines may be classified as .81 (81%) Topic 1 ("vaccines"), .10 (10%) Topic 2 ("government"), etc., to a sum of 1 (i.e., documents have some proportion of *all* topics, but usually fall into one or two topics, based on the parameter alpha (Section 4.1.3.2)). One of the foundations of *LDA* is the *Dirichlet Distribution*, a "distribution of distributions" modeled by several parameters outlined below. For a more thorough description of *LDA*, see Blei, Ng, and Jordan (2003).

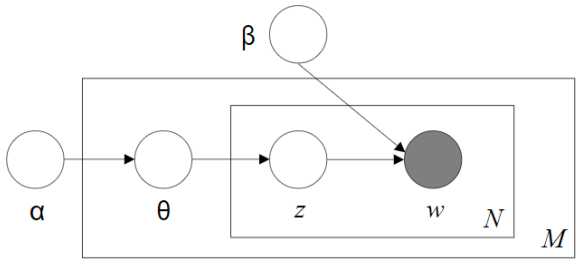

Figure 1: *Graphical model representation of LDA. From Blei (2003).*

Initially, the *LDA* model was performed on a sample of ≈ 20,000 Tweets, because the size and resource requirements of *LDA* grow exponentially as new documents are added to the model (Blei, Ng, and Jordan 2003). This sample was able to provide adequate coverage of the original dataset, with the topics of the remainder of the corpus being predicted using the sample *LDA* model as training data. The parameters used to control the *LDA* model, as well as those generated as an output, are described below:

4.1.3.1 K The parameter k defines the number of discrete topics modeled using unsupervised classification. In this study, a value of $k = 11$ was chosen by iterating from $k = 2$ to $k = 24$ and calculating the *coherence score* of each value of k – the coherence score measures the similarity of terms within each topic, and is a rough measure of how well the *LDA* model assigns topics to each term and document. The *short-text* nature of Twitter data provides several complica-

tions when performing topic modeling; one such issue was the generally low coherence values (≈ 0.1). As such, the value of k was also chosen to minimize overlap between distinct topics, rather than focus solely on the maximum coherence score. Even when optimizing topic coherence, the value of k is somewhat arbitrary, with the final decision being left to the researchers' discretion (Alghamdi and Alfalqi 2015).

4.1.3.2 Alpha (α) The α parameter models the distribution of topics within each document; at low values of α (i.e., close to 0), documents are likely to consist primarily of only one topic. At values of $\alpha \approx 1$, all distributions of topics per document are equally likely. As $\alpha \to \infty$, all topics become equally likely to occur (i.e., with $k = 3$ topics, documents are composed of 33% topic 1, 33% topic 2, and 33% topic 3). α is controlled by a single argument within the `FitLdaModel` function, with a value of 0.1.

4.1.3.3 Beta (β) The β parameter effectively models the distribution of words per topic; as β decreases, topics are composed of few terms, while high values of β generate topics with larger numbers of terms. The value of β for this analysis was set to 0.05. For this analysis, a smaller value of β (< 1) provided topics with little overlap between terms. One limitation of *short-text* Twitter data is that documents often have only a few dozen words, making traditional *LDA* more challenging than when performed e.g. on book chapters or customer reviews. Changing the values of β to include more terms might lead to greater overlap between topics, but also to topics with generally higher coherence.

4.1.3.4 Phi (ϕ) The output parameter ϕ models the probability of tokens (words) falling into each topic; i.e., for each topic, the probabilities of each token falling into that topic. ϕ is used to model the topics themselves to find the terms that are most frequently used within each topic, in order to gain an intuitive understanding of what each topic represents semantically. Table 1 shows the top 5 terms within each topic, which were calculated using the values of ϕ (Jones 2019).

4.1.3.5 Theta (θ) The value θ of *LDA* is a matrix which represents the distribution of topics over documents; that is, it is the output which is used to assign topics to each document. Based on the Dirichlet distribution defined in part by α, the values across rows of θ sum to 1; each document is made of a certain proportion of every topic, with most documents consisting of one "primary" topic. For the sake of simplicity when performing multinomial logistic regression (see Section 4.2.2, the maximum value of θ within each document was used to assign only a single topic to each document; further research involving specific, topic-wise values of θ could be done to provide a continuous dependent and independent variable within regression analysis.

Further research involving the optimal parameters of α, β, and k is warranted; given constraints regarding time and computational resources, the values for these parameters were not optimized, remaining at their default values assigned by the `textmineR::FitLdaModel` function (Jones 2019).

4.2 Analysis

4.2.1 Topic Name Assignment.

Table 1: Top 5 terms (ϕ) within each topic

Topic	Terms
Big-Pharma	people, stop, cdc, fuck, shit
COVID-19-Outbreaks	covid, coronavirus, virus, cdc, amp
COVID-19-Testing	test, covid, testing, people, tested
Government-Skepticism	cdc, trump, people, trust, science
Masks	mask, vaccinated, masks, people, wear
Public-Health	health, amp, covid, care, public
Quarantine-Distancing	kids, school, schools, home, children
Sanitation	masks, face, mask, virus, hands
U.S.-Cases-Deaths	covid, cases, deaths, numbers, states
Vaccine-Side-Effects	vaccine, covid, people, vaccines, shot
Vaccine-Skepticism	immunity, covid, vaccine, long, natural

A descriptive name for each topic was generated with the `textmineR::SummarizeTopics` function,

which automatically assigns each topic a label based on most prevalent terms. The outcome of this function was then "cleaned up" and given proper capitalization and punctuation for legibility. This function was used to aid in eliminating potential researcher bias in arbitrarily assigning names to topics.

When generating a Document-Term Matrix using the `textmineR::CreateDtm` function, the argument `doc_names` is used to link `tweet_id` to the output of the *LDA* model (i.e., the rows of the matrix θ). Topics were assigned to documents using the maximum value of θ within each row; further research should use the entire distribution of θ to use a set of continuous variables within multinomial logistic regression. The resulting table of topics was then joined to the original corpus to create a data frame of Tweets containing text, `tweet_id`, `conversation_id`, and topic.

Note should be taken regarding the topics "Big-Pharma", "Government-Skepticism", and "Vaccine-Skepticism", specifically in how these topics are applied to CDC Tweets. As the CDC does not intentionally promote skepticism towards the efficacy of mask-wearing and vaccines, these topics warrant further examination. CDC Tweets of these topics are primarily artefacts of how *LDA* was used in this analysis; these are largely Tweets that either do not clearly fall into topics regarding public health and mask-wearing, or were authored pre-COVID, when much of the discourse consisted of regulatory information ("vaping," food-related recalls, etc.).

In order to set up regression analysis, Tweets had to be "linked" by conversation ID. For this analysis, CDC Tweets were set to be the "parent" of the conversation, and the corpus of Reply Tweets was filtered and grouped within each `conversation_id`. The resulting table contained, for each distinct `conversation_id`, the parent CDC Topic as well as a collection of any Reply Tweets and their Topics. This allowed for a direct relationship between CDC Tweets and Reply Tweets, so that their respective topics could be compared using regression.

4.2.2 Multinomial Logistic Regression

The primary method of regression analysis for this study was *multinomial logistic regression*, a method capable of modeling the predicted response of a categorical dependent variable with more than two possible outcomes (i.e. non-binary) (Ripley and Venables 2021). Multinomial logistic regression was used over other regression models such as ordinal logistic regression, as the response variable was not in any particular order (e.g., a Likert scale), meaning the data was nominal.

Using the `multinom` function from the `nnet` R package, multinomial logistic regression was performed on the corpus. Data were aggregated by CDC Tweets using conversation ID, such that Reply Tweets of the same conversation ID were in the same group. Multinomial logistic regression compares values of response variables to changes from a "baseline" value, in this study the `Public-Health` topic, such that a change in the predictor variable leads to an expected change in the response variable.

5 Results

5.1 Topic Modeling

The results from topic modeling through *LDA* (see Table 1) indicate that the primary discussion topics relating to the CDC include mask wearing (`Masks-Mask-Efficacy`), vaccine efficacy and skepticism (`Vaccine-Side-Effects` and `Vaccine-Skepticism`). As the results from *MLR* indicate (below), the most "controversial" CDC Topic, defined as the topic which generated the widest array of skeptical Reply Topics, was `Masks-Mask-Efficacy`, indicating a strong public mistrust in government mask mandates. The number of topics chosen for *LDA*, \$k =\$11, was chosen to provide adequate coverage of all topics of discussion,

while also limiting the amount of overlap between distinct topics.

5.2 Multinomial Logistic Regression

Table 3 (see Appendix) gives the results of multinomial logistic regression in the form of Odds Ratios (OR), which indicate the relative "odds" of the occurrence of a Reply Topic per CDC Topic, when compared to the baseline of "Public Health". As this baseline was chosen arbitrarily, some consideration should be given as to what these Odds Ratios actually represent. For example, in Column 2 (the CDC Topic *COVID-19-Outbreaks*), a value of 2.11 is given in Row 9 (Reply Topic *Vaccine-Side-Effects*); this represents that, with a CDC Tweet of the topic `Covid-19-Outbreaks`, Reply Tweets are 2.11 times more likely to be of the topic `Vaccine-Side-Effects` than of the topic `Public-Health`.

The CDC Topics which are deemed the most controversial are those which generate the widest variety of Reply Topics; interpreting the Odds Ratios in Table 3, these Topics are those with several values greater than 1.

6 Discussion

6.1 Controversial Topics

The Topics generated by *Latent Dirichlet Allocation* show a pattern of discussion among users that indicates a strong mistrust in government programs regarding mask wearing, vaccine mandates, public health, and basic sanitation. The most "controversial" of these is the `Masks-Mask-Efficacy` topic, which generated the largest and most divisive conversations; much of this division is linked to changes in federal mask mandates, specifically around the January 20, 2021 executive order requiring mask wearing in federal buildings, and in March 8, 2021, when the CDC no longer required fully vaccinated Americans to social distance or wear masks indoors (Netburn 2021).

Understanding the controversy behind these topics is important in the field of science communication, as it allows for clearer messaging to individuals and communities of various educational and socioeconomic backgrounds, so that public health campaigns remain inclusive and available for all.

6.2 Applications

Several useful potential applications for this research exist; first, by using topic modeling, a better understanding of public perception of government programs can be achieved by "boiling down" potentially millions of data points into a discrete set of topics. Second, the relationship between Reply Topics and Parent (CDC) Topics can be modeled through logistic regression to establish a direct connection between the semantic topic of a public health message and the topics of user discussion. This relationship is vital to understand as it allows for better tailoring of public health messaging and science communication in general. By using machine learning models such as *LDA*, organizations can predict the expected Topic of a given message, which then can be used in conjunction with a logistic regression model to predict the expected distribution of topics of replies.

7 Conclusion

Twitter is a valuable source of information regarding sentiment and perception of public health organizations, policies, and messages, and offers direct access into the discussions made by individuals affected by such policies. In order to better understand these discussions, this study performed text mining analysis using topic modeling to attain relationships between public health messages and user viewpoints. Topic modeling, performed through the use of *Latent Dirichlet Allocation*, provides a significant insight into the abstract discussion topics acquired from

social media data. Tweets were assigned topics using the ϕ and θ parameters, which allowed for linking topics to conversations defined by each conversation's unique `conversation_id`. The results of this operation were conversations containing the Parent (CDC) Topic and all replies of varying Reply Topics. Multinomial Logistic Regression was then performed to model the causal relationship between a CDC Tweet and its Reply Tweets. The regression model shows how each CDC Topic affects the probability and distribution of the occurrences of each Reply Topic.

References

Alghamdi, R., and K. Alfalqi. 2015. A survey of topic modeling in text mining. *Int. J. Adv. Comput. Sci. Appl.(IJACSA)* 6 (1).

Bao, S., S. Xu, L. Zhang, R. Yan, Z. Su, D. Han, and Y. Yu. 2009. Joint emotion-topic modeling for social affective text mining. In *2009 Ninth IEEE International Conference on Data Mining*, 699–704. IEEE.

Barrie, C., and J. Ho. 2021. *academictwitteR: An R package to access the Twitter Academic Research Product Track v2 API endpoint.*

Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research* 3:993–1022.

Boon-Itt, S., and Y. Skunkan. 2020. Public perception of the COVID-19 pandemic on Twitter: Sentiment analysis and topic modeling study. *JMIR Public Health and Surveillance* 6 (4):e21978.

Bouchet-Valat, M. 2020. *SnowballC: Snowball stemmers based on the c 'libstemmer' UTF-8 library.*

Choli, M., and D. J. Kuss. 2021. Perceptions of blame on social media during the coronavirus pandemic. *Computers in Human Behavior* :106895.

Clavel, C., and Z. Callejas. 2015. Sentiment analysis: From opinion mining to human-agent interaction. *IEEE Transactions on affective computing* 7 (1):74–93.

Debortoli, S., O. Müller, I. Junglas, and J. vom Brocke. 2016. Text mining for information systems researchers: An annotated topic modeling tutorial. *Communications of the Association for Information Systems* 39 (1):7.

Dubey, A. D. 2020. Twitter Sentiment Analysis during COVID-19 Outbreak. *Available at SSRN 3572023*.

Garcia, K., and L. Berton. 2021. Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA. *Applied Soft Computing* 101:107057.

Hong, L., and B. D. Davison. 2010. Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics*, 80–88.

Huang, J., and R. Liu. 2020. Xenophobia in America in the Age of Coronavirus and Beyond. *Journal of Vascular and Interventional Radiology* 31 (7):1187.

Jones, T. 2019. *textmineR: Functions for Text Mining and Topic Modeling.*

Liu, B. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies* 5 (1):1–167.

Manguri, K. H., R. N. Ramadhan, and P. R. M. Amin. 2020. Twitter sentiment analysis on worldwide COVID-19 outbreaks. *Kurdistan Journal of Applied Research* :54–65.

Muñoz, A., C. Moreno, and J. L. Luján. 2012. Who is willing to pay for science? On the rela-

tionship between public perception of science and the attitude to public funding of science. *Public Understanding of Science* 21 (2):242–253.

Negara, E. S., D. Triadi, and R. Andryani. 2019. Topic Modelling Twitter Data with Latent Dirichlet Allocation Method. In *2019 International Conference on Electrical Engineering and Computer Science (ICECOS)*, 386–390. IEEE.

Netburn, D. 2021. A timeline of the CDC's advice on face masks. *Los Angeles Times*.

Ostrowski, D. A. 2015. Using latent Dirichlet allocation for topic modelling in twitter. In *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)*, 493–497. IEEE.

Pak, A., and P. Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, 1320–1326.

Rashid, J., S. M. A. Shah, and A. Irtaza. 2019. Fuzzy topic modeling approach for text mining over short text. *Information Processing & Management* 56 (6):102060.

Ripley, B., and W. Venables. 2021. *Nnet: Software for feed-forward neural networks with a single hidden layer, and for multinomial log-linear models.*

Twitter API Documentation. 2021. *Twitter Developer Platform*.

Zulfikar, W. A., and V. Auliansyah. 2020. Public perception of physical distancing in preventing the spread of coronavirus disease (COVID-19) in the city of Banda Aceh in 2020. *EXECUTIVE EDITOR* 11 (7):1579.

8 Appendix

The following appendix provides much of the R code used in this study.

8.1 Tables and Figures

Table 2: Example of Tweets aggregated by conversation ID

conversation_id	label_CDC	label_Reply	Count
1440058230260916237	Vaccine-Side-Effects	Quarantine-Self-Isolation	9
1440058230260916237	Vaccine-Side-Effects	U.S.-Cases-Deaths	9
1440058230260916237	Vaccine-Side-Effects	Public-Health	4
1440058230260916237	Vaccine-Side-Effects	COVID-19-Outbreaks	9
1440058230260916237	Vaccine-Side-Effects	Handwashing-Sanitation	2
1440058230260916237	Vaccine-Side-Effects	Government-Skepticism	4
1440058230260916237	Vaccine-Side-Effects	Vaccine-Skepticism	14
1440058230260916237	Vaccine-Side-Effects	Vaccine-Side-Effects	33
1440058230260916237	Vaccine-Side-Effects	COVID-19-Testing-Symptoms	7
1440058230260916237	Vaccine-Side-Effects	Big-Pharma	28
1440058230260916237	Vaccine-Side-Effects	Masks-Mask-Efficacy	11

Table 3: Odds Ratios: Reply Topic per CDC Topic

	(Intercept)	Big-Pharma	COVID-19-Outbreaks	COVID-19-Testing-Symptoms	Government-Skepticism	Handwashing-Sanitation	Masks-Mask-Efficacy	Quarantine-Self-Isolation	U.S.-Cases-Deaths	Vaccine-Side-Effects	Vaccine-Skepticism
Big-Pharma	1.24	2.99	1.15	1.27	1.35	1.80	2.54	2.29	1.83	2.07	1.65
COVID-19-Outbreaks	0.69	1.36	1.90	1.52	1.60	2.00	1.52	1.78	1.62	1.46	1.84
COVID-19-Testing-Symptoms	0.76	0.62	1.48	4.39	0.86	1.33	1.33	1.19	1.51	1.26	1.33
Government-Skepticism	1.26	4.65	1.20	1.09	1.74	1.81	1.65	1.58	1.87	1.40	1.04
Handwashing-Sanitation	0.60	1.01	1.49	1.41	0.95	6.47	2.15	2.27	1.64	1.30	1.21
Masks-Mask-Efficacy	2.18	0.40	0.50	0.57	0.42	0.99	1.58	0.89	0.97	0.73	0.63
Quarantine-Self-Isolation	0.67	1.14	1.40	1.17	1.36	2.06	2.06	5.55	2.21	1.81	1.13
U.S.-Cases-Deaths	0.72	0.79	1.56	1.64	0.86	1.39	2.02	1.52	4.21	1.67	1.64
Vaccine-Side-Effects	0.67	0.94	2.11	1.96	0.65	1.33	2.64	2.30	1.85	6.13	5.30
Vaccine-Skepticism	0.71	1.00	1.47	1.87	1.41	1.43	3.48	1.55	1.80	4.26	3.69

[1] CDC Topics by Column; Reply Topics by Row

[2] Note: 'Public-Health' is absent as it is the baseline for MLR analysis.

8.2 academictwitteR

```
date.start <- "2019-01-01T00:00:00Z"
date.end <- "2021-09-21T00:00:00Z"

get_all_tweets(
  users = c("CDCgov"),
  start_tweets = date.start,
  end_tweets = date.end,
  n = Inf,
  data_path = "CDCtweets/",
  bind_tweets = F
)

get_all_tweets(
  query = "@CDCgov",
  start_tweets = date.start,
  end_tweets = date.end,
  n = Inf,
  data_path = "ReplyTweets/",
  bind_tweets = F
)
```

8.3 Latent Dirichlet Allocation

```
library(textmineR)

# Read in Tweets acquired from academictwitteR::get_all_tweets()


tweets_full <- read_csv(
  "tweets_full.csv",
  col_types = cols(tweet_id = col_character(),
                   conversation_id = col_character())
)

# Filter Tweets by English language and select only necessary columns

tweets_filter <- tweets_full %>%
  filter(lang == "en") %>%
  select(
    tweet_id, conversation_id, text
  )

# Listing Twitter-specific stopwords
t.stopwords <- c("amp", "$", "RT")

# Create a Document-Term Matrix of 1- and 2-grams
# Also, this step removes stopwords and lemmaticizes the data
tweets_DTM <- CreateDtm(
    doc_vec = tweets_filter$text,
    doc_id = tweets_filter$tweet_id,
    ngram_window = c(1,2),
    stopword_vec = c(
      stopwords::stopwords(language = "en"),
      stopwords::stopwords(source = "smart"),
      t.stopwords
    ),
    stem_lemma_function = function(x){
      SnowballC::wordStem(words = x, language = "porter")
    },
    remove_punctuation = T,
    remove_numbers = T,
    # Convert to lower-case
    lower = T
  )


# Finally, generating the LDA Model

tweets_LDA <- FitLdaModel(
  dtm = tweets_DTM,
  k = 11,
  iterations = 1000,
  burnin = 500,
  calc_coherence = T,

  # These parameters should be optimized in the publication run; this model
  # takes a lot of time to process so the values are kept at default.
  alpha = .1,
  beta = .05
)

# Create a dataframe based on the theta parameter by tweet_id
theta <- as.data.frame(tweets_LDA$theta) %>%
  rownames_to_column("tweet_id")


# Joining theta to tweets_filter gives us a distribution of topics per Tweet
tweets.LDA <- inner_join(tweets_filter, theta)


# This function allows for a programmatic labeling of each topic, rather than try
# and label these manually (and induce bias!)
LDA.summary <- SummarizeTopics(LDA)



# Making some of the topics a bit more legible; the names are still based off of
# the top terms generated by `SummarizeTopics`
topic.names <- tibble(
  labels = c(
    "american_people",
    "big_pharma",
    "covid_-",
    "covid_vaccine",
    "natural_immunity",
    "public_health",
    "stay_home",
    "tested_positive",
    "united_states",
    "wash_hands",
    "wear_mask"
  ),
  names = c(
    "Government-Skepticism",
    "Big-Pharma",
    "COVID-19-Outbreaks",
    "Vaccine-Side-Effects",
    "Vaccine-Skepticism",
    "Public-Health",
    "Distancing",
    "COVID-19-Testing",
    "U.S.-Cases-Deaths",
    "Sanitation",
    "Masks"
  )
)
topic.names.c <-
  setNames(as.character(topic.names$names),
           topic.names$labels)


# Recoding the LDA Summary to include the more legible names
LDA.summary <- LDA.summary %>%
  dplyr::mutate(
    Topic_Name = recode(label_1,
                        !!!topic.names.c) %>%
      factor(ordered = F)
  ) %>% arrange(Topic_Name)

# Just a simple table of topic and topic names, for recoding later data
topic.description <- LDA.summary %>%
  select(
    topic, Topic_Name
  )

# This function calculates the most significant topic by finding the maximum
# value of theta for each Tweet.
tweets.LDA$max.topic <-
  colnames(tweets.LDA[, 4:14])[max.col(tweets.LDA[, 4:14])]

# Same as above, but returns the actual value of theta; not really used in the
# analysis, but good to see just for reference
tweets.LDA$max.theta <- apply(tweets.LDA[, 4:14], 1, max)


# Applying the LDA model to Tweets made by the CDC; this is a major shortcoming
# of this analysis, and is something I'd like to fix before publication. Some
# changes include actually generating a *separate* LDA model unique to CDC
# Tweets, as the topics they discuss are very different than those of the
# general population.

# Loading the .csv with readr...
```

```
tweets_CDC <- readr::read_csv("tweets_cdc.csv") %>%
  dplyr::select(tweet_id, text, conversation_id)

# Fitting the LDA model requires a DTM
CDC_DTM <- CreateDtm(
  doc_vec = tweets_CDC$text,
  doc_names = tweets_CDC$tweet_id,
  ngram_window = c(1, 2),
  stopword_vec = c(
    stopwords::stopwords(language = "en"),
    stopwords::stopwords(source = "smart"),
    t.stopwords
  ),
  stem_lemma_function = function(x) {
    SnowballC::wordStem(words = x, language = "porter")
  },
  remove_punctuation = T,
  remove_numbers = T,
  lower = T
)


# Predicting CDC Topics based on original LDA; again, this is something that
# will change in the final version of this paper.
CDC_topics <- predict(object = tweets_LDA,
                      newdata = CDC_DTM)

# For brevity, I won't include all the following steps, because they're
# essentially the same as what was done for the Reply Tweets
```

```
# Tweets were grouped by conversation_id. Unfortunately, the original file
# containing the actual analysis has been lost, but the process of grouping
# by conversation_id is roughly the same as follows:


# "CDC_topics" and "Reply_topics" are dataframes of tweet_id,
# conversation_id, and topic
tweets_conversation <- CDC_topics %>%
  left_join(Reply_topics, by = "conversation_id")

# Below is an example of what the grouped dataframe looks like for a single
# conversation_id. The "count" column indicated how many Reply Tweets of a
# Reply Topic are in that conversation
```

8.4 Multinomial Logistic Regression

```
mlr <- conversation_long %>%
  dplyr::filter(
    # Filtered out rows where no Reply Tweets of a topic were found.
    value > 0
  ) %>%
  multinom(
    # label_CDC (CDC Topic) is the only independent variable in this
    # case, but the results are weighted by `value`, or the count of
    # Reply Tweets of different Topics (Count column in the example
    # table above)
    formula = label_Reply ~ label_CDC,
    weights = value,
    data = .)

# Calculate the Odds Ratio from the Coefficients from MLR.
odds <- exp(coef(mlr)) %>%
  as.data.frame()
```