Topic Modeling of Twitter Data regarding the CDC and the COVID-19 Pandemic

Harrison Brown

2021-11-22

Contents

1 Abstract

Social media is a powerful source of data regarding individual perception of public health policy and phenomena such as the COVID-19 pandemic. In this study, we perform topic modeling – a method of determining abstract "topics" within collections of documents – on approximately 300,000 Tweets from January 1, 2019, to September 21, 2021, to better understand both user sentiment and conversational topics within the discourse. Topic modeling is performed on Tweets by the CDC's official Twitter account *"@CDCgov"*, as well as Tweets that reply to, quote, or mention the CDC's Twitter handle. Topic modeling is performed using *Latent Dirichlet Allocation (LDA)*, with a value of $k = 11$ topics. The results of the multinomial logistic regression show the odds ratio of Reply Tweets of a given topic occurring on CDC Tweets of each topic, indicating that certain, potentially "controversial" tweets by the CDC , such as those regarding mask mandates and vaccinations, are significantly likely to produce replies of specific, "skeptical" topics, e.g. discussions of "Big Pharma" and government skepticism.

expand abstract?

2 Introduction

2.1 Research Question

This paper aims to develop the foundation of a method for predicting discussion topics on social media using *Latent Dirichlet Allocation (LDA)* and multinomial logistic regression (MLR), for the purposes of increasing effectiveness of science communication programs. As much of the literature focuses on the societal effects of the pandemic *itself*, this study aims to measure user opinion from the perspective of public health geography by examining the relationship between the topics of Tweets by the CDC (*CDC Topic* and *CDC Tweet*) and their effects on the topics of Tweets made in reply (*Reply Topic* and *Reply Tweet*). This study intends to answer the following research question; fundamentally, how can social media managers (whether for governmental or non-governmental organizations) use topic modeling to better understand public opinion to tailor more effective public health campaigns.

3 Literature Review

While much research regarding text mining of Twitter data has focused specifically on the COVID-19 pandemic, little has been done to examine 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).

3.1 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 the data in this analysis consists of *short-text,* with many Tweets containing fewer than 10 words), sentiment analysis was not performed within this study, the focus instead placed on abstract connections between topics generated by the methodology described below in Section 5.1.3.

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.

3.2 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)..

For this study, much of the literature involved 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).

4 Data

4.1 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 n.d.). 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.

4.2 academictwitteR

Query by "@CDCgov" for Reply Tweets, and by "from:CDCgov" for CDC Tweets.

Barrie and Ho (2021)

Table 1: Top 5 terms within each topic

Topic	Top Terms (ϕ)
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

4.3 Data Preprocessing

4.3.1 Stopword Removal

4.3.2 Duplicates / Noise

4.3.3 Conversation ID

4.3.4 Tokenization

4.3.5 Stemming

5 Methodology

5.1 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 5.1.3 for more detail.

5.1.1 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 4.3. 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.

5.1.2 Document-Term Matrix

5.1.3 Latent Dirichlet Allocation

Latent Dirichlet Allocation (hereafter *LDA*), developed in Blei, Ng, and Jordan (2003), is an unsupervised machine-learning approach to topic modeling, in which topics are assigned through "fuzzy clustering" into different subsets of topics. 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 (α)).

One of the foundations of *LDA* is the *Dirichlet Distribution*, a "distribution of distribution" modeled by several parameters outlined below. For a more thorough description of *LDA*, see Blei, Ng, and Jordan (2003).

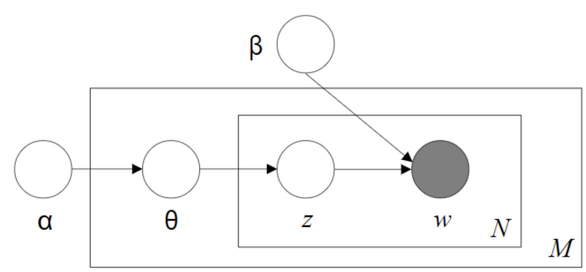

Figure 1: *Graphical model representation of LDA. The boxes are "plates" representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document. From Blei (2003).*

5.1.3.1 ***K*** The parameter K, or 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.

5.1.3.2 Alpha (α)

The α parameter models the distribution of topics within each document; at low values of α (i.e., close to 0), topics 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).

5.1.3.3 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.

5.1.3.4 Phi (ϕ) and Gamma (γ)

5.1.3.5 Theta (θ)

5.1.3.6 *Iterations*

5.2 Analysis

5.2.1 Topic Name Assignment.

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 purposes. This function was used to aid in eliminating potential researcher bias in arbitrarily assigning names to topics.

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.).

5.2.2 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).

Using the `multinom` function from the `nnet` R package, multinomial logistic regression was calculated on each *conversation id*. **...explain `conversation_long` etc....** This function requires a "baseline" explanatory and response variable, which was chosen to be the "Public-Health" topic.

6 Results

Table 2: Odds Ratios: Reply Topic per CDC Topic

	(Intercept)	Big-Pharma	COVID-19-Outbreaks	COVID-19-Testing-Symptoms	Government-Skepticism
Big-Pharma	1.24	2.99	1.15	1.27	1.35
COVID-19-Outbreaks	0.69	1.36	1.90	1.52	1.60
COVID-19-Testing-Symptoms	0.76	0.62	1.48	4.39	0.86
Government-Skepticism	1.26	4.65	1.20	1.09	1.74
Handwashing-Sanitation	0.60	1.01	1.49	1.41	0.95
Masks-Mask-Efficacy	2.18	0.40	0.50	0.57	0.42
Quarantine-Self-Isolation	0.67	1.14	1.40	1.17	1.36
U.S.-Cases-Deaths	0.72	0.79	1.56	1.64	0.86
Vaccine-Side-Effects	0.67	0.94	2.11	1.96	0.65
Vaccine-Skepticism	0.71	1.00	1.47	1.87	1.41

	Handwashing-Sanitation	Masks-Mask-Efficacy	Quarantine-Self-Isolation	U.S.-Cases-Deaths	Vaccine-Side-Effects	Vaccine-Skepticism
Big-Pharma	1.80	2.54	2.29	1.83	2.07	1.65
COVID-19-Outbreaks	2.00	1.52	1.78	1.62	1.46	1.84
COVID-19-Testing-Symptoms	1.33	1.33	1.19	1.51	1.26	1.33
Government-Skepticism	1.81	1.65	1.58	1.87	1.40	1.04
Handwashing-Sanitation	6.47	2.15	2.27	1.64	1.30	1.21
Masks-Mask-Efficacy	0.99	1.58	0.89	0.97	0.73	0.63
Quarantine-Self-Isolation	2.06	2.06	5.55	2.21	1.81	1.13
U.S.-Cases-Deaths	1.39	2.02	1.52	4.21	1.67	1.64
Vaccine-Side-Effects	1.33	2.64	2.30	1.85	6.13	5.30
Vaccine-Skepticism	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.

7 Discussion

8 Conclusion

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. *Journal of Open Source Software* 6 (62):3272.

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.

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.

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.

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.

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.

Twitter API Documentation.