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

Harrison Brown

2021-11-21

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 probability of Reply Tweets of a given topic occurring on CDC Tweets of each topic, indicating that certain, potentially "controversial" tweets by the CDC are significantly likely to produce replies of specific, "skeptical" topics.

expand abstract?

Contents

List of Figures

List of Tables

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.

2.2 Background and Lit. Findings

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).

2.2.1 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 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 @ref(latent-dirichlet-allocation).

3 Data

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 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.

3.2 academictwitteR

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

Barrie and Ho (2021)

3.3 Data Preprocessing

3.3.1 Stopword Removal

3.3.2 Duplicates / Noise

3.3.3 Conversation ID

3.3.4 Tokenization

3.3.5 Stemming

4 Methodology

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 @ref(tab:topicsummary) 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 @ref(latent-dirichlet-allocation) for more detail.

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

Terms joined by '_' represent bigrams.

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 @ref(data-preprocessing). 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

4.1.3 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:

4.1.3.1 *K*

4.1.3.2 Alpha (α)

4.1.3.3 Beta (β)

4.1.3.4 Phi (ϕ) and Gamma (γ)

4.1.3.5 Theta (θ)

4.1.3.6 *Iterations*

4.2 Analysis

4.2.1 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.

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).

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.

5 Results

Table 2: 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.

6 Discussion

7 Conclusion

References

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. https://joss.theoj.org/papers/10.21105/joss.03272.

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):7493.

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. https://linkinghub.elsevier.com/retrieve/pii/S1568494620309959.

Liu, B. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies* 5 (1):1167.

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* :5465.

Pak, A., and P. Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. 13201326.

Twitter API Documentation. https://developer.twitter.com/en/docs/twitter-api.