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

Harrison Brown

2021-11-19

## 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 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.

# 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).

# 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 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.

Table 1: Top 5 terms within each topic

| Topic | Top 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 |

Terms joined by '_' represent bigrams.

#### 4.1.1 Constructing Corpus

#### 4.1.2 Document-Term Matrix

#### 4.1.3 Latent Dirichlet Allocation

Blei, Ng, and Jordan (2003)

## 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-Outbreaks | COVID-Testing | Government-Skeptic | Sanitation | Masks | Distancing | US-Cases | Vaccine-Effects | Vaccine-Skeptic |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Big-Pharma | 1.2437 | 2.9938 | 1.1510 | 1.2727 | 1.3465 | 1.8028 | 2.5408 | 2.2889 | 1.8317 | 2.0728 | 1.6484 |
| COVID-19-Outbreaks | 0.6885 | 1.3572 | 1.8967 | 1.5236 | 1.5997 | 1.9974 | 1.5219 | 1.7774 | 1.6150 | 1.4554 | 1.8437 |
| COVID-19-Testing-Symptoms | 0.7647 | 0.6172 | 1.4786 | 4.3946 | 0.8624 | 1.3340 | 1.3349 | 1.1869 | 1.5104 | 1.2575 | 1.3333 |
| Government-Skepticism | 1.2560 | 4.6488 | 1.1979 | 1.0938 | 1.7359 | 1.8142 | 1.6536 | 1.5849 | 1.8660 | 1.4049 | 1.0398 |
| Handwashing-Sanitation | 0.6019 | 1.0096 | 1.4879 | 1.4066 | 0.9545 | 6.4653 | 2.1521 | 2.2730 | 1.6358 | 1.3007 | 1.2076 |
| Masks-Mask-Efficacy | 2.1842 | 0.3995 | 0.4990 | 0.5721 | 0.4161 | 0.9910 | 1.5782 | 0.8881 | 0.9739 | 0.7342 | 0.6279 |
| Quarantine-Self-Isolation | 0.6703 | 1.1400 | 1.4039 | 1.1684 | 1.3608 | 2.0602 | 2.0612 | 5.5542 | 2.2120 | 1.8137 | 1.1295 |
| U.S.-Cases-Deaths | 0.7200 | 0.7916 | 1.5591 | 1.6385 | 0.8605 | 1.3917 | 2.0199 | 1.5222 | 4.2104 | 1.6677 | 1.6367 |
| Vaccine-Side-Effects | 0.6664 | 0.9442 | 2.1067 | 1.9628 | 0.6454 | 1.3291 | 2.6361 | 2.3018 | 1.8462 | 6.1258 | 5.2994 |
| Vaccine-Skepticism | 0.7106 | 0.9969 | 1.4657 | 1.8667 | 1.4148 | 1.4307 | 3.4831 | 1.5503 | 1.7975 | 4.2610 | 3.6903 |

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

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.

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.

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