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

Harrison Brown

2021-11-18

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; see Table 1.

2 Introduction

2.1 Research Question

2.2 Background and Lit. Findings

3 Data

3.1 Twitter API

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

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

5 Results

Table 1: Odds Ratios: Reply Topic per CDC Topic

	Intercept	Big-Pharma	COVID-Outbreaks	COVID-Testing	Government-Skeptic	Sanitation	Masks	Distancing	US-C
Big-Pharma	1.2437	2.9938	1.1510	1.2727	1.3465	1.8028	2.5408	2.2889	1.8
COVID-19-Outbreaks	0.6885	1.3572	1.8967	1.5236	1.5997	1.9974	1.5219	1.7774	1.6
COVID-19-Testing-Symptoms	0.7647	0.6172	1.4786	4.3946	0.8624	1.3340	1.3349	1.1869	1.5
Government-Skepticism	1.2560	4.6488	1.1979	1.0938	1.7359	1.8142	1.6536	1.5849	1.8
Handwashing-Sanitation	0.6019	1.0096	1.4879	1.4066	0.9545	6.4653	2.1521	2.2730	1.6
Masks-Mask-Efficacy	2.1842	0.3995	0.4990	0.5721	0.4161	0.9910	1.5782	0.8881	0.9
Quarantine-Self-Isolation	0.6703	1.1400	1.4039	1.1684	1.3608	2.0602	2.0612	5.5542	2.2
U.S.-Cases-Deaths	0.7200	0.7916	1.5591	1.6385	0.8605	1.3917	2.0199	1.5222	4.2
Vaccine-Side-Effects	0.6664	0.9442	2.1067	1.9628	0.6454	1.3291	2.6361	2.3018	1.8
Vaccine-Skepticism	0.7106	0.9969	1.4657	1.8667	1.4148	1.4307	3.4831	1.5503	1.7

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.