Government Health Communication During the COVID-19 Pandemic: A BERT Topic Modeling Approach

Thomas Moore-Pizon
University of South Florida,
Tampa, FL, USA
temoore@usf.edu

Nic DePaula
SUNY Polytechnic Institute
Utica, NY, USA
nfvd@proton.me

Loni Hagen
University of South Florida,
Tampa, FL, USA
lonihagen@usf.edu

Abstract

Different levels of government agencies have exerted great effort to communicate with the public during the Covid-19 pandemic on multiple social media platforms. This study uses BERT topic modeling, an artificial intelligence model, to extract topics from various public health agencies of cities, states and the federal government on Twitter and Facebook for the years 2020 and 2021. We contrast and compare major topics addressed by these agencies related to Covid-19 and the pandemic across the two major social media platforms. The findings show how we can employ BERT topic modeling to extract social media topics during a health emergency and evaluate the extent to which topics covered by these agencies address the major social and health concerns of the pandemic.

Introduction

Government agencies have an important role in communicating health information to the public on social media. Social media are important sites of public communication, and people expect reliable, relevant, and timely information from government agencies, especially in the context of a health emergency (CDC, 2018). During the Covid-19 pandemic, government agencies were active on social media platforms, and so was the public engaging with the content (Sutton et al, 2020). However, while many studies have examined characteristics of government health communication during the COVID-19 pandemic, few studies have examined how to employ automated topic modeling techniques to understand the content of this communication over time, and the various topics addressed by government agencies through multiple years of the pandemic.

In this study, we employ BERTopic to analyze government health communication topics on social media and show differences and similarities in topics covered across agencies and platforms. BERTopic provides a state-of-the-art approach to discovering topics in a corpus of documents by identifying topic-revealing terms in the corpus. This approach is useful as it enables semi-automated analyses of large datasets from various organizations over a long time (e.g. multiple years). Secondly, it enables interactive visualizations of content that may be embedded in intelligent systems, such as dashboards for tracking government coverage of a health emergency or similar situation. Lastly, this approach provides an additional way for social scientists, policymakers, and the public to understand the topics being addressed by public health departments across regions and waves of a crisis, enabling us to evaluate how well government agencies are addressing the related issues.

Related Work

Many studies have adopted machine learning models to track public health discourse on social media (Argyris et al., 2021; Benis et al., 2021). However, previous studies have not employed computational *topic modeling* for long-term analysis of topics by government agencies during the Covid-19 pandemic. Topic modeling based on Latent Dirichlet Allocation (LDA) has been widely adopted for extracting naturally arising themes in large volumes of text data (Hagen, 2018; Hagen et al., 2016). Despite their popularity, some of the major weaknesses of LDA methods were in 1) pre-defining the number of topics to produce, 2) removing stop words (insignificant terms), and 3) their inability to properly reflect semantics and contexts of documents (Angelov, 2020; Egger & Yu, 2022).

To overcome these weaknesses, BERT topic modeling was developed. BERT (Bidirectional Encoder Representations

from Transformers), introduced by Google (Ravichandiran, 2021), is a language model based on the transformer architecture. BERTopic utilizes transformer-based BERT for document embedding generation, which produces better quality topics that can reflect semantics, contexts, and nuances expressed in text data (Xie et al., 2023; Egger & Yu, 2022). Using pre-trained language model BERT, BERTopic first converts each document to its embedding representation, then reduces the dimensionality of the embeddings, which is followed by clustering. From these clustered documents, topic representations are extracted.

Methodology

Data Collection

Datasets for this study were collected in late 2021 (Facebook data) and late 2023 (Twitter data), We first identified United States (US) federal agencies related to health and infectious disease, all US state public health agencies, and several public health agencies for the largest cities in the US that had a Facebook or Twitter account. We then retrieved all Facebook posts from these agencies' accounts from January 2020 to July 2021 (for Facebook), and from January 2020 to December 2021 (for Twitter) based on availability of data and restrictions from the APIs. We then retrieved all posts (not replies or retweets) referring to Covid-19 based on the presence of either keyword: 'ncov', 'covid', 'corona', 'pandemic', or 'sars-cov'. Ultimately, the dataset contains 99,794 Twitter Covid-related posts and 77,228 Facebook Covid-related posts from 172 distinct accounts.

Methods

We used the BERTopic method by Grootendorst (2022) to extract topics from our datasets. The only text pre-processing applied was to remove URLs due to the extreme noises it brought to the results. For this study, we set the default parameters suggested by Grootendort (2022), and the topic numbers were set to "auto," which produced over 100 topics. Among the produced topics, we analyzed the top 20 topics, based on the number of documents reflecting the topic. We created a graph with dates in x-axis and number of documents representing the topic in y-axis for each of the six datasets (three levels of government per platform) for the comparison.

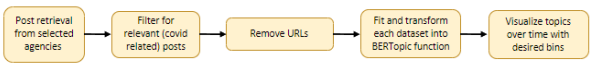

Figure 1: Workflow for Methods

Results

Our results show coherent topics from our model, distinct trends of over time, and differences and similarities across agency levels and platforms, which we visualize in several ways but only summarize here due to page restrictions.

Federal Health Communication

In both platforms, Federal agencies communicated about 1) *preventive* measures such as washing hands, maintaining 6 feet apart, wearing masks, and 2) *risk factors* for severe illness. Some major differences were that on Twitter we observed, among others, topics on 1) *stress and anxiety* and 2) *delta variant*, which was initially mentioned in mid-2021, when it was becoming prevalent. On Facebook agencies addressed 1) *school related* topics, mainly in the beginning of the pandemic, 2) *vaccine safety*, and 3) *general media briefs*.

State Health Communication

State agencies communicated, among other topics, 1) *official announcements*, 2) *updates* on cases, 3) *nurse appreciation*, and 4) how to deal with *stress and anxiety* on both Facebook and Twitter. At the state level, we also observed a focus on the *delta variant* and references to *contact tracing*, while this did not appear on Facebook major topics.

Local Health Communication

Local agencies addressed some of the same issues already mentioned (e.g. *wearing masks* and *preventive measures*) and included references to local issues, such as *closed hours of test sites, town hall questions* and *levels of hospitalizations*. However, *town hall questions* were a topic observed on Facebook but not on Twitter. We also observed on Twitter that *children* become a prevalent topic on local Twitter accounts, mostly related to children vaccine, but this topic did not appear on local Facebook topics.

Limitation and Future Directions

The results of this analysis show a potential for BERTopic to extract coherent topics from government health communication on social media. However, limitations of this analysis include the outsized representation of certain accounts which posted more than others (e.g. a *New York* topic appears often). Although we used internal validation methods to identify the optimum number of topics, it was not clear these results provided the best result set. In future work, the analysis may balance the amount of content across agencies, and test different methods for topic size selection. Moreover, a method for validating the representativeness of the topics will be necessary. Lastly, we plan to experiment with improving accuracy by adopting a COVID-Twitter_BERT (Müller et al., 2023), a specialized model for Twitter data.

Acknowledgment

This work was partially supported with a grant from the Florida Center for Cybersecurity.

References

Angelov, D. (2020). Top2Vec: Distributed Representations of Topics (arXiv:2008.09470). arXiv. https://doi.org/10.48550/arXiv.2008.09470

Argyris, Y. A., Monu, K., Tan, P.-N., Aarts, C., Jiang, F., & Wiseley, K. A. (2021). Using Machine Learning to Compare Provaccine and Antivaccine Discourse Among the Public on Social Media: Algorithm Development Study. JMIR Public Health and Surveillance, 7(6), e23105. https://doi.org/10.2196/23105

Benis, A., Chatsubi, A., Levner, E., & Ashkenazi, S. (2021). Change in Threads on Twitter Regarding Influenza, Vaccines, and Vaccination During the COVID-19 Pandemic: Artificial Intelligence–Based Infodemiology Study. JMIR Infodemiology, 1(1), e31983. https://doi.org/10.2196/31983

CDC (2018). Crisis and Emergency Risk Communication (CERC): Introduction. *Center for Disease Control and Prevention.*

Egger, R., & Yu, J. (2022). A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology*, *7*. https://doi.org/10.3389/fsoc.2022.886498

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure (arXiv:2203.05794). arXiv. https://doi.org/10.48550/arXiv.2203.05794

Hagen, L. (2018). Content analysis of e-petitions with topic modeling: How to train and evaluate LDA models? Information Processing & Management, 54(6), 1292–1307. https://doi.org/10.1016/j.ipm.2018.05.006

Hagen, L., Harrison, T. M., Uzuner, Ö., May, W., Fake, T., & Katragadda, S. (2016). E-petition popularity: Do linguistic and semantic factors matter? Government Information Quarterly, 33(4), 783–795. https://doi.org/10.1016/j.giq.2016.07.006

Müller, M., Salathé, M., & Kummervold, P. E. (2023). COVID-Twitter-BERT: A natural language processing model to analyse COVID-19 content on Twitter. Frontiers in Artificial Intelligence, 6, 1023281. https://doi.org/10.3389/frai.2023.1023281

Ravichandiran, S. (2021). Getting Started with Google BERT: Build and train state-of-the-art natural language processing models using BERT. Packt Publishing Ltd.

Sutton, J., Renshaw, S. and Butts, C. (2020). COVID-19: Retransmission of Official Communications in an Emerging Pandemic. *PLOS ONE,* 15(9). e0238491.

Xie, T., Ge, Y., Xu, Q., & Chen, S. (2023). Public Awareness and Sentiment Analysis of COVID-Related Discussions Using BERT-Based Infoveillance. AI, 4(1), Article 1. https://doi.org/10.3390/ai4010016