Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)

doi:10.3233/SHTI230266

Investigating Canadian Public Attitudes Toward COVID-19 Vaccine Mandates with a Nested Analysis Framework

Yang YANG[a,1], Xiaomeng LEI[b], Zeping ZHOU[b], Shu-Feng TSAO[a], Zahid A. BUTT[a] and Helen H. CHEN[a]
[a]*School of Public Health Sciences, Canada*
[b]*Computer Science, University of Waterloo, Canada*

Abstract. Background: Social media is an important medium for studying public attitudes toward COVID-19 vaccine mandates in Canada, and Reddit network communities are a good source for this. **Methods**: This study applied a "nested analysis" framework. We collected 20378 Reddit comments via the Pushshift API and developed a BERT-based binary classification model to screen for relevance to COVID-19 vaccine mandates. We then used a Guided Latent Dirichlet Allocation (LDA) model on relevant comments to extract key topics and assign each comment to its most relevant topic. **Results**: There were 3179 (15.6%) relevant and 17199 (84.4%) irrelevant comments. Our BERT-based model achieved 91% accuracy trained with 300 Reddit comments after 60 epochs. The Guided LDA model had an optimal coherence score of 0.471 with four topics: travel, government, certification, and institutions. Human evaluation of the Guided LDA model showed an 83% accuracy in assigning samples to their topic groups. **Conclusion**: We develop a screening tool for filtering and analyzing Reddit comments on COVID-19 vaccine mandates through topic modelling. Future research could develop more effective seed word-choosing and evaluation methods to reduce the need for human judgment.

Keywords. COVID-19, Canada, vaccine mandate, public attitudes, Reddit, social media, topic modelling, BERT, Guided LDA

1. Introduction

Social media data has been used to study public health issues and attitudes toward vaccination [1]. Reddit is an important communication channel, and its volume of data grows with web technology advances. To analyze large amounts of text data, it is necessary to develop methods for classifying and filtering this "Big Data," often using NLP techniques. Zhu *et al.* [2] developed a pre-training method based on the Bidirectional Encoder Representations from Transformers (BERT) model to classify Tweets about personal experiences with medication. Klein *et al.* [3] proposed using the BERT model to identify potential COVID-19 cases by collecting and filtering tweets with self-reported symptoms and geolocations. Muller, Salathé, and Kummervold [4] developed CT-BERT, a BERT model optimized for COVID-19 tasks by being pre-trained on Twitter data related to the pandemic.

[1] Corresponding Author: Yang Yang, E-mail: y24yang@uwaterloo.ca

Latent Dirichlet Allocation (LDA) is a probabilistic model for learning the structure of discrete data collections, such as corpora [5,6]. It is frequently used to discover topics and analyze social media data. LDA has been applied to Twitter data, achieving an accuracy of 98% in classifying topics [7]. Researchers have used LDA to analyze the themes and patterns present in tweets and examine the relationships between those themes [8]. The Guided LDA model is a variant of LDA that addresses limitations such as topic overlapping and meaningless topics. It allows the specification of seed words to guide the model to focus on certain terms and has an *eta* parameter that tracks word assignment to topics. However, using a prior distribution with seed words can influence the model to concentrate on specific topics. This project aimed to analyze Canadian public discourse about COVID-19 vaccine mandate policies announced in September 2021 by applying a "nested analysis" framework and topic modelling techniques to Reddit data collected in Canada.

2. Methods

2.1. Related Work

Researchers have proposed various pre-training and fine-tuning methods based on BERT and RoBERT models [9] for COVID-19-related tasks. Sentence-BERT (SBERT) [10] was introduced to compare semantically meaningful sentence embeddings, and BioBERT [11] was proposed for analyzing biomedical documents. Latent Semantic Analysis (LSA) [12] and LDA are mathematical methods that have been applied to topic modelling and sentiment analysis of Twitter data. LDA has shown better accuracy than LSA in topic modelling [7] and is useful in identifying tweet patterns, themes, and structures [8].

2.2. Dataset and Preprocessing

We collected 20378 Reddit comments from Canada in September 2021 and labelled each comment as relevant or irrelevant to COVID-19 vaccine mandates. Of the total samples, 17199 (84.4%) comments were irrelevant, and 3179 (15.6%) were relevant comments. An under-sampling strategy was employed to create a balanced training set α with an equal amount of 3179 relevant and irrelevant comments. The data preprocessing steps comprised word tokenization, stopword removal, word stemming, and synonyms merging.

2.3. Binary Classification Using BERT-based Model

The BERT-based binary classifier with DistilRoBERTa was trained using different amounts of data, ranging between 100 and 1000 comments randomly selected from α. The BERT-based model maps sentences to a 768-dimensional vector space, producing a feature vector for each comment and using a neural network with two fully connected layers and an output layer with Relu, Relu, and Softmax activation functions for binary classification.

2.4. Topic Modelling Using Guided LDA

We initially tried the standard LDA model to explore vaccine mandate topics, but it struggled to split themes. We then developed a Guided LDA model, which selects seed words related to each topic based on the standard LDA model results to improve performance.

3. Results

3.1. Binary Classification Using BERT-based Model

The BERT-based model performed best with training on 300 comments randomly selected from α, achieving 91% accuracy on the entire dataset after 60 epochs (Table 1).

Table 1. BERT-based binary classification model performance with different sizes of datasets.

Size of Dataset (# of Reddit comments)	Accuracy	Convergence Speed (Epochs)
100	0.83	275
150	0.86	150
200	0.89	100
300	**0.91**	60
400	0.90	60
500	0.90	40
1000	0.89	25

The loss of the training set decreased from 0.7 to 0.3, and the accuracy increased to approximately 0.9 (Figure 1).

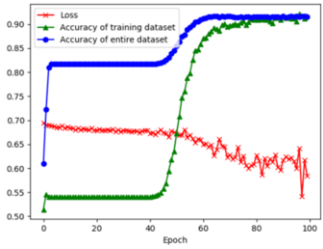

Figure 1. The BERT-based binary classification model trained on 300 Reddit comments.

3.2. Topic Modelling Using Guided LDA

The Guided LDA model divided the Reddit comments into four topics based on the optimal coherence score of 0.47 (Figure 2). The model's performance was evaluated by manual interpretation, which showed that 83% of the samples were accurately assigned to their topic group. Table 2 summarizes the four topics and corresponding seed words.

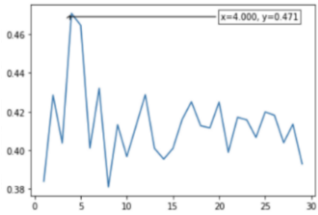

Figure 2. Coherence scores of Guided LDA model for different topic numbers. X-axis: number of topics; Y-axis: coherence score

Table 2. Seed words for Guided LDA.

Topic	Seed Words
Travel	"test" "travel" "day" "border" "result" "hour" "quarantine" "pcr" "time" "flight" "trip" "plan" "home" "country" "week" "return" "airport" "enter" "hotel"
Government	"govern" "law" "right" "province" "policy" "court" "gov" "protest" "employ" "vote" "leadership" "party" "elect" "win" "province" "mask" "rule" "service"
Certification	"code" "card" "qr" "passport" "health" "system" "app" "id" "proof" "receipt" "mandate" "certificate" "restrict"
Institution	"work" "employer" "employee" "hospital" "business" "school" "kid" "number" "case" "rate"

4. Discussion

This study aimed to understand the public attitudes toward COVID-19 vaccine mandates in Canada by analyzing comments on the Reddit platform using a nested analysis framework. We developed a pipeline that utilizes a BERT-based model for binary classification of Reddit comments related to COVID-19 vaccine mandates and a Guided LDA model to further analyze the topics in the relevant comments. Our study found that Canadian public discourse on COVID-19 vaccine mandates revolved around four main themes: travel, government, certification, and institutions. These themes encompassed discussions on the impact of vaccine requirements on travel [13], the government's role in implementing and enforcing mandates [14,15], proof of vaccination and vaccine passports [16], and the role of schools, workplaces, health care, and other institutions in implementing mandates [14,15].

The BERT-based model achieved an accuracy of 91% with a small training set of 300 Reddit comments and converged faster with larger training sets, demonstrating its usefulness in identifying relevant Reddit comments related to COVID-19 vaccine mandates. Our results also showed that having more training data did not always lead to better performance, as the model's accuracy dropped slightly to 89% when trained on 1000 comments. The Guided LDA model achieved 83% accuracy in human evaluation, improving standard LDA through careful tuning with seed words for each topic. This allowed for a more accurate understanding of public attitudes toward COVID-19 vaccine mandates from a large volume of Reddit comments. Still, the quality of the seed words and evaluation limitations may affect the model's performance and require manual interpretation.

This study applies a "nested analysis" framework with NLP topic modelling and human judgement for model evaluation and iterative data preprocessing. The Pushshift API allowed for efficient analysis of large quantities of Reddit data over a specific time period with greater flexibility and access to additional parameters, making it superior to keyword searches with the Reddit search engine. This approach allowed for discovering emerging topics and identifying patterns and themes using topic modelling techniques such as LDA. One limitation is that Reddit comments may not represent the overall Canadian population's views on COVID-19 vaccine mandates and may not be reliable sources of information. Future research should consider alternative methods for classifying datasets, evaluating the Guided LDA model, and selecting seed words to improve the model's performance.

5. Conclusion

We use NLP techniques on Reddit comments to investigate Canadian attitudes toward COVID-19 vaccine mandates and demonstrate their usefulness for health researchers. Filtering social media data and focusing on specific aspects, especially when dealing with large amounts of text data, can improve understanding of public health issues and inform effective interventions.

References

[1] Müller MM, Salathé M. Crowdbreaks: tracking health trends using public social media data and crowdsourcing. Frontiers in public health. 2019 Apr 12;7:81.
[2] Zhu M, Song Y, Jin G, Jiang K. Identifying personal experience tweets of medication effects using pre-trained Roberta language model and its updating. In Proceedings of the 11th international workshop on health text mining and information analysis 2020 Nov (pp. 127-137).
[3] Klein AZ, Magge A, O'Connor K, Flores Amaro JI, Weissenbacher D, Gonzalez Hernandez G. Toward using Twitter for tracking COVID-19: a natural language processing pipeline and exploratory data set. Journal of medical Internet research. 2021 Jan 22;23(1):e25314.
[4] Müller M, Salathé M, Kummervold PE. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. arXiv preprint arXiv:2005.07503. 2020 May 15.
[5] Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000 Jun 1;155(2):945-59.
[6] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. Journal of machine Learning research. 2003;3(Jan):993-1022.
[7] Negara ES, Triadi D, Andryani R. Topic modelling twitter data with latent dirichlet allocation method. In2019 International Conference on Electrical Engineering and Computer Science (ICECOS) 2019 Oct 2 (pp. 386-390). IEEE.
[8] Xue J, Chen J, Chen C, Zheng C, Li S, Zhu T. Public discourse and sentiment during the COVID 19 pandemic: Using Latent Dirichlet Allocation for topic modeling on Twitter. PloS one. 2020 Sep 25;15(9):e0239441.
[9] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692. 2019 Jul 26.
[10] Reimers N, Gurevych I. Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084. 2019 Aug 27.
[11] Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020 Feb 15;36(4):1234-40.
[12] Landauer TK, Foltz PW, Laham D. An introduction to latent semantic analysis. Discourse processes. 1998 Jan 1;25(2-3):259-84.
[13] Charrier L, Garlasco J, Thomas R, Gardois P, Bo M, Zotti CM. An Overview of Strategies to Improve Vaccination Compliance before and during the COVID-19 Pandemic. International Journal of Environmental Research and Public Health. 2022 Sep 3;19(17):11044.
[14] Griffith J, Marani H, Monkman H. COVID-19 vaccine hesitancy in Canada: content analysis of tweets using the theoretical domains framework. Journal of medical Internet research. 2021 Apr 13;23(4):e26874.
[15] Benham JL, Atabati O, Oxoby RJ, Mourali M, Shaffer B, Sheikh H, Boucher JC, Constantinescu C, Parsons Leigh J, Ivers NM, Ratzan SC. COVID-19 vaccine–related attitudes and beliefs in Canada: National cross-sectional survey and cluster analysis. JMIR Public Health and Surveillance. 2021 Dec 23;7(12):e30424.
[16] Flores Anato JL, Ma H, Hamilton MA, Xia Y, Harper S, Buckeridge D, Brisson M, Hillmer MP, Malikov KA, Kerem A, Beall R. Impact of a vaccine passport on first-dose COVID-19 vaccine coverage by age and area-level social determinants in the Canadian provinces of Québec and Ontario: an interrupted time series analysis. medRxiv. 2022:2022-10.