BERT-Powered Multi-label Classifier: Analyzing Public COVID Vaccination Discourse

Ranjit Patro[1], Asutosh Mishra[2]

[1]*Indian Institute of Science Education and Research Berhampur, Odisha, India*
[2]*Indian Institute of Science Education and Research Berhampur, Odisha, India*

Abstract

In response to the pressing need to understand public sentiment against vaccination in the digital era, this study employs social media data to build a multi-label, multi-class classifier. To accomplish effective label prediction, a pre-trained BERT model is used in conjunction with certain preprocessing techniques, and a brute-force threshold selection strategy is implemented. This research sheds light on the complex terrain of vaccination opinion by analyzing a wide range of concerns, ranging from the safety of vaccines to their potential political and religious ramifications. This study is part of the Artificial Intelligence on Social Media (AISoMe) track of the Forum for Information Retrieval Evaluation (FIRE) 2023 conference. The evaluation of the test dataset shows that the test score of 0.7 is meaningful.

Keywords

Multi-Label Classifier, COVID-19 Vaccine Tweets, COVID-Twitter-BERT

1. Introduction

Vaccination has historically served as a fundamental pillar of public health, providing protection to communities against the grave threats posed by deadly diseases. In the present-day globalized society, the discussion pertaining to vaccines has gained significant traction on social media platforms. The COVID-19 pandemic has presented unique and unparalleled challenges, leading to increased attention on the crucial importance of immunization in protecting public health. Nevertheless, the current period is characterized not alone by significant scientific progress, but also by an increasing sense of doubt regarding vaccines.

The discourse surrounding vaccines on social media encompasses a wide range of perspectives and concerns, resulting in a complex and multifaceted narrative. Utilizing data from social media platforms, this study intends to investigate the complex landscape. Our primary objective is to develop a robust multi-label, multi-class classifier capable of effectively categorizing social media conversations, specifically tweets, by accurately identifying the specific vaccine-related concerns mentioned by the authors. The study also examines the complex nature of vaccine skepticism, which encompasses various aspects such as concerns regarding effectiveness, safety, the role of the pharmaceutical sector, and the broader socio-political and cultural factors.

Forum for Information Retrieval Evaluation, December 15-18, 2023, India

✉ ranjitp20@iiserbpr.ac.in (R. Patro); asutosh21@iiserbpr.ac.in (A. Mishra)
🌐 https://ranjitpatro.netlify.com/ (R. Patro); https://asutosh-mishra.netlify.com/ (A. Mishra)
iD 0009-0006-4701-7098 (R. Patro); 0009-0003-5376-3431 (A. Mishra)

CEUR Workshop Proceedings (CEUR-WS.org)

The value of this work extends beyond its timely reaction to the current vaccine controversy, as it also holds the potential to provide valuable insights for informing public health measures. Through an in-depth analysis of the intricate concerns and beliefs that underpin vaccine reluctance, this study provides valuable insights about vaccination sentiment. This technology facilitates the understanding and analysis of public opinion towards vaccines in the era of digital communication, so promoting the development of more informed public health.

2. Task

Under the Artificial Intelligence on Social Media (AISoMe) track [1], in this paper, we introduce an effective approach to address the challenge of constructing a robust multi-label, multi-class classifier for categorizing social media posts, specifically tweets, based on the various concerns expressed by the authors regarding vaccines. Our classification task involves assigning the most appropriate label(s) to each tweet from a set of potential concerns associated with vaccines. These concerns encompass a wide spectrum of opinions and viewpoints in the discourse around vaccines, acting as the label of classification. They are:

- **Unnecessary**: The tweet indicates vaccines are unnecessary, or that alternate cures are better.
- **Mandatory**: Against mandatory vaccination — The tweet suggests that vaccines should not be made mandatory.
- **Pharma**: Against Big Pharma — The tweet indicates that the Big Pharmaceutical companies are just trying to earn money, or the tweet is against such companies in general because of their history.
- **Conspiracy**: Deeper Conspiracy — The tweet suggests some deeper conspiracy, and not just that Big Pharma wants to make money (e.g. vaccines are being used to track people, COVID is a hoax)
- **Political**: Political side of vaccines — The tweet expresses concerns that the governments or politicians are pushing their own agenda through vaccines.
- **Country**: Country of origin — The tweet is against some vaccine because of the country where it was developed or manufactured.
- **Rushed**: Untested or Rushed Process — The tweet expresses concerns that the vaccines have not been tested properly or that the published data is not accurate.
- **Ingredients**: Vaccine Ingredients or technology — The tweet expresses concerns about the ingredients present in the vaccines (e.g. fetal cells, chemicals) or the technology used (e.g. mRNA vaccines can change your DNA)
- **Side-effect**: Side Effects or Deaths — The tweet expresses concerns about the side effects of the vaccines, including deaths caused.
- **Ineffective**: Vaccine is Ineffective — The tweet expresses concerns that vaccines are not effective enough and are useless.
- **Religious**: Religious Reasons — The tweet is against vaccines because of religious reasons
- **None**: No specific reason stated in the tweet, or some reason other than the given ones.

A single tweet can encompass one or multiple distinct concerns regarding vaccine viewpoints, as demonstrated by the following examples:

- “It begins. Please find safe alternatives to this vaccine. UK issues allergy warning about Pfizer COVID-19 vaccine after patients fall ill https://t.co/JEHgCLGIbv via @nypost”; **Labels: side-effect**
- “@BorisJohnson @CMO_England @MattHancock THIS IS BEYOND INCOMPETENCE People should refuse this vaccine as it’s not going to be administered correctly. Only a matter of time before Covid19 kills one of your patients who have had only a fraction of intended protection!”; **Labels: none**
- “Dare I suggest something more sinister with Johnson suggesting a vaccine passport aimed at the reopening of pubs would force the young to seek out the vaccine. A vaccine that is experimental offers the individual no protection from getting or passing on the virus”; **Labels: mandatory, ineffective**
- “jeffmcnamee @padakitty @Amanda77197114 @alexanderchee BREAKING: FDA announces 2 deaths of Pfizer vaccine trial participants from “serious adverse events. textquotedbl Fed Up Democrats Say NO to Forced Vaccines in NY”; **Labels: side-effect, mandatory, political**

3. Related Work

Today, users of microblogs such as Twitter contribute a wide variety of content, including their ideas and feelings in relation to topics such as the coronavirus, COVID-19 immunizations, and vaccination campaigns. Extracting meaningful information from textual tweets has become an integral aspect of social computing. Text classification in particular has been successful through the use of a variety of methods, which range from more conventional machine learning techniques such as Naive-Bayes, Linear classifiers, and Support Vector Machines to more advanced deep learning approaches such as Long Short Term Memory (LSTM) and Bidirectional Recurrent Neural Networks. In addition, modern language models, such as **BERT** (Bidirectional Encoder Representations from Transformers) [2], as well as its domain-specific variations, such as CT-BERT (COVID-Twitter-BERT) [3], and the improved CT-BERT-V2 (COVID-Twitter-BERT-V2), exemplify the cutting-edge advancements in natural language processing.

3.1. BERT

The bidirectional contextual comprehension of BERT (Bidirectional Encoder Representations from Transformers) allows it to analyze and capture the nuanced details in textual input, making it a key Natural Language Processing (NLP) paradigm. Pre-trained on vast unlabeled text corpora and subjected to pre-training tasks like Masked Language Modeling and Next Sentence Prediction, BERT acquires profound linguistic understanding. Furthermore, the effectiveness of this approach is emphasized by its flexibility to be easily adjusted for specific NLP downstream tasks by incorporating layers tailored to those tasks. This consistently leads to achieving the best results in a wide range of applications.

4. Dataset

In this study, we utilize a comprehensive train and test dataset provided under the AISoME track. This meticulously curated train dataset encompasses a corpus of 9,921 anti-vaccine tweets focused on COVID vaccines, originally posted on twitter during the period spanning 2020-21. [4] Notably, each of these tweets has been diligently annotated by human experts, associating them with one or more of the aforementioned labels to facilitate fine-grained analysis. Subsequently, the train dataset comprises the annotated tweets, accompanied by their corresponding tweet IDs and assigned labels. It is imperative to note that a single tweet may exhibit multiple labels, reflecting the multifaceted nature of vaccine-related concerns. Similarly, The test dataset featuring 486 tweets, furnished with tweet IDs, which, while unlabelled, encompass discussions concerning a spectrum of vaccines, extending beyond COVID vaccines to encompass other vaccine types such as the MMR vaccine and the Flu vaccine. This whole dataset serves as the foundation for our research endeavors, enabling nuanced exploration and analysis of diverse vaccine-related concerns.

5. Pre-processing

The tweets within the provided dataset exhibit a diverse range of unique lexicons, including elements such as '@username' mentions, http-urls, hashtags, and special characters like emojis. While these elements may convey contextual information in certain contexts, they introduce noise into the dataset, potentially hindering the overall performance of the model, particularly in the context of our study. To ensure the integrity and effectiveness of our analysis, we devised a systematic data-cleaning pipeline as an integral component of our pre-processing procedure. This pipeline entails a series of procedural steps designed to enhance the quality of the dataset by mitigating the impact of extraneous elements. These steps include:

- **Conversion to Lowercase**: We first convert all sentences to lowercase. This uniform casing ensures that the analysis operates consistently on words and phrases, effectively reducing potential inconsistencies arising from varying letter cases.
- **Removal of Non-Alphanumeric Characters**: Leveraging Python's regular expression library, we eliminate all non-alphanumeric characters from the text. This step serves to mitigate noise introduced by non-textual elements.
- **Elimination of URLs**: Given the presence of URLs in the raw tweet data, we employ regular expressions to systematically remove them. Specifically, any word matching the pattern beginning with "http" and followed by one or more non-whitespace characters is excised, effectively eradicating URLs from the dataset.
- **Exclusion of Usernames**: Twitter utilizes the '@username' format to mention specific individuals within tweets. Recognizing this, we remove all usernames by identifying words containing the character '@' and subsequently omitting them from the text.
- **Stopword Removal**: To prioritize meaningful information, we also implement the removal of stopwords—commonly occurring words like "the," "a," "an," and "in"—that typically contribute limited substantive value to the analysis.

6. Methodology

6.1. Model

CT-BERT-V2 is a transformer-based language model for Twitter discourse analysis during the COVID-19 pandemic. It stands out for its domain-specific focus, having been pre-trained on a substantial corpus of tweets containing relevant keywords such as "wuhan", "ncov", "coronavirus", "covid" and "sars-cov-2" posted from January 12 to July 5, 2020. CT-BERT-V2 is initialized with BERT-Large and fine-tuned on over 97 million tweets and 1.2 billion training samples. This customised method allows CT-BERT-V2 to decipher COVID-related Twitter discussions, making it a powerful tool for assessing pandemic data. Now, for our work using the given dataset, to improve model robustness and mitigate overfitting, we introduce a dropout layer with a 0.5 probability atop the pre-trained CT-BERT-V2. Subsequently, a linear transformation layer is incorporated, mapping the 1024-dimensional embeddings from BERT-Large to a final 11-dimensional output logit. These logits form the basis for loss calculation and are further processed through a sigmoid activation function, yielding the ultimate output probabilities for all 11 labels excluding the "None" label.

6.2. Experimental Setup

In our experimental setup, after pre-processing, we employed a crucial data transformation step involving the use of the MultiLabelBinarizer provided by sklearn, which facilitated the process of binarizing the training data through One Hot Encoding Method. During this transformation, we excluded the 'none' column, recognizing it as a dependent label only present when all other labels are absent. This deliberate approach, known as handling the dummy variable trap, was essential in ensuring the accuracy of our results. As a result, the final dataset consisted of 11 labels, with labels represented as "none" when the corresponding tweet provides a value of 0 for all other labels. Following this transformation, we meticulously partitioned the dataset into distinct training and validation sets, while preserving the distribution of class instances within each set and maintaining a balanced representation.

For the fine-tuning of CT-BERT-V2, we exclusively employed the pre-processed training data, while the validation data played a pivotal role in evaluating model performance. To calculate the loss, we utilized the logits generated by the model, applying the Binary Cross-Entropy with Logits loss function. These logits were then subjected to a sigmoid function, producing 11 normalized values, each corresponding to the probability of a specific label as output. This well-defined experimental pipeline served as the foundation for our comprehensive evaluation and analysis[1].

6.3. Prediction

In this phase, we faced the challenge of predicting tweets having multiple labels. To fix this, we used a threshold mechanism to return predictions for all 11 labels with probabilities above a certain threshold. The 'none' label was chosen when none of the 11 labels exceeded this criterion.

[1] GitHub link to the work: `https://github.com/Ranjit246/AISoME_FIRE_2023`

To optimize this threshold-setting process effectively, we adopted a greedy approach. Recognizing the variations in sigmoid output due to the presence of imbalanced training data, we chose to employ distinct thresholds for each of the 11 labels. This necessitated the generation of threshold values through an exhaustive exploration of 100 threshold values per label, generated using numpy linspace within the range of 0 to 1. The threshold selected for each label was determined based on the maximum macro-F1 score achieved for that specific label during the assessment on a randomized validation split. This iterative process was replicated across all 11 labels, resulting in a comprehensive list of 11 "best thresholds". Subsequently, these thresholds underwent validation on different splits and were further fine-tuned to yield the final list of threshold values. Ultimately, classes with probabilities greater than or equal to their corresponding class threshold were identified as the final predicted classes, ensuring a precise and robust multi-label classification outcome.

7. Evaluation

The assessment of results in the AISoMe Track is conducted using the widely adopted standard classification metric, the Macro-F1 score, encompassing 12 distinct classes. In Table 1, we present the outcome of our submission for the task. Remarkably, our prepared model achieved a Macro-F1 score of 0.7 and an equally commendable Jaccard score of 0.71. This stellar performance underscores the effectiveness and competence of our approach in addressing the complex challenges posed by multi-label classification in the context of the AISoMe Track.

Table 1
Team IISERBPR-NLP, Result of AISoME track

Run File	Summary of Methodology	Macro-F1	Jaccard	Rank
submission-bert.csv	fine tune BERT with best threshold	0.7	0.71	3

8. Conclusion and Future Work

In our study, We used Covid-Twitter-BERT, a transformer-based model pre-trained on a large corpus of COVID-19-related tweets, to efficiently assign vaccine-related labels for a challenging multi-label and multi-class task. Transformer-based models are data-hungry, thus we plan to investigate data augmentation ways to improve our model. We also consider using adversarial training to strengthen our model. Our approach will be improved by these future research paths, allowing us to better analyze vaccine concerns in social media discussions. In addition, we plan to expand our dataset sources to include more social media platforms, optimize our model through fine-tuning and hyperparameter tuning, prioritize model explainability, consider real-time monitoring capabilities, address ethical concerns, and foster collaborative partnerships for domain-specific insights to enable practical vaccine applications.

References

[1] S. Poddar, M. Basu, K. Ghosh, S. Ghosh, Overview of the fire 2023 track:artificial intelligence on social media (aisome), in: Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation, 2023.

[2] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805.

[3] M. Müller, M. Salathé, P. E. Kummervold, Covid-twitter-bert: A natural language processing model to analyse COVID-19 content on twitter, CoRR abs/2005.07503 (2020). URL: https://arxiv.org/abs/2005.07503.

[4] S. Poddar, A. M. Samad, R. Mukherjee, N. Ganguly, S. Ghosh, Caves: A dataset to facilitate explainable classification and summarization of concerns towards covid vaccines, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 3154–3164.