

Enhancing Multilabel Classification of Anti-Vaccine Tweets with the COVID-Twitter-BERT

Aritra Mandal¹

¹A.K Chowdhury School of Information Technology, University of Calcutta, Kolkata, West Bengal, India

Abstract

Social media platforms have revolutionized global communication, enabling billions of individuals to share their perspectives and opinions. With the worldwide rollout of COVID-19 vaccination campaigns, the classification of anti-vaccine tweets assumes significance as it offers valuable insights into people's concerns about the new vaccines. These tweets not only provide feedback but also offer a glimpse into the specific apprehensions people hold, ranging from potential side effects to concerns related to vaccine effectiveness and political influences. In this research, we employ a specialized BERT model tailored to the domain, achieving notable performance with a macro-F1 score of 0.71 and a Jaccard score of 0.72.

Keywords

BERT, anti-vaccine tweets, classification

1. Introduction

In the face of the COVID-19 pandemic, the world finds itself engaged in one of the most formidable battles in recent history. Historically, vaccines have emerged as a reliable and effective weapon against infectious diseases, conferring immunity to individuals and contributing to the global efforts to combat and eradicate deadly viruses. The rapid development and distribution of COVID-19 vaccines have exemplified the power of science and international collaboration, offering hope for a return to normalcy.

Amidst the vaccine rollout, an unprecedented dialogue has unfolded across social media platforms, with Twitter emerging as a prominent arena for discussions surrounding COVID-19 vaccines. These discussions encompass a wide spectrum of topics, ranging from the progress of vaccination campaigns, accessibility issues, and vaccine efficacy to the possible side effects. Within this digital discourse, a diverse array of opinions prevails, spanning from enthusiastic support to pronounced skepticism.

Recognizing the significance of these online conversations, government entities, health organizations such as the World Health Organization (WHO), and public health experts have a vested interest in understanding public sentiment and concerns regarding the new COVID-19 vaccines. The insights drawn from these micro-blogs offer invaluable guidance for shaping future strategies to promote widespread vaccination. As such, a crucial aspect of this endeavor


FIRE'23: Forum for Information Retrieval Evaluation, December 15-18, 2023, India

✉ aritramandal37@gmail.com (A. Mandal)

🆔 0009-0008-1841-5120 (A. Mandal)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

involves the thorough analysis and interpretation of the reasons underlying vaccine hesitancy and resistance.

2. Task

The goal is to build an effective multi-label classifier to label a social media post (particularly, a tweet) according to the specific concern(s) towards vaccines as expressed by the author of the post.[1] Note that a tweet can have more than one label (concern), e.g., a tweet expressing 3 different concerns towards vaccines will have 3 labels. We consider the following concerns towards vaccines as the labels for the classification task:

1. Unnecessary: The tweet indicates vaccines are unnecessary, or that alternate cures are better.
2. Mandatory: Against mandatory vaccination – The tweet suggests that vaccines should not be made mandatory.
3. 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.
4. Conspiracy: Deeper Conspiracy – The tweet suggests some deeper conspiracy, and not just that the Big Pharma want to make money (e.g., vaccines are being used to track people, COVID is a hoax)
5. Political: Political side of vaccines – The tweet expresses concerns that the governments/politicians are pushing their own agenda though the vaccines.
6. Country: Country of origin – The tweet is against some vaccine because of the country where it was developed / manufactured
7. Rushed: Untested / Rushed Process – The tweet expresses concerns that the vaccines have not been tested properly or that the published data is not accurate.
8. Ingredients: Vaccine Ingredients/technology – The tweet expresses concerns about the ingredients present in the vaccines (eg. fetal cells, chemicals) or the technology used (e.g., mRNA vaccines can change your DNA)
9. Side-effect: Side Effects / Deaths – The tweet expresses concerns about the side effects of the vaccines, including deaths caused.
10. Ineffective: Vaccine is ineffective – The tweet expresses concerns that the vaccines are not effective enough and are useless.
11. Religious: Religious Reasons – The tweet is against vaccines because of religious reasons
12. None: No specific reason stated in the tweet, or some reason other than the given ones.

3. Related Work

Due to the exponential growth of social media platforms, sharing content on these platforms has expanded tremendously, further increasing the malicious content on these platforms[2][3]. Therefore detection of such malicious content has gained significant attraction among the research community. The traditional machine learning methods like Naive-Bayes classifier, Linear

classifier, Support Vector Machine and Deep neural methods like Long Short Term Memory (LSTMs) and Bidirectional RNN are very successful for text classification. More recent language models for natural language processing include BERT (Bidirectional Encoder Representations from Transformers)[4] and its domain-specific version CT-BERT (COVID-Twitter-BERT) [5].

3.1. BERT

BERT (Bidirectional Encoder Representations from Transformers) revolutionizes natural language processing by capturing contextual relationships in both directions of a sequence. Introduced by Google in 2018, BERT employs a transformer architecture, considering the entire input context to enhance understanding and generate more accurate language representations. Its pre-training on massive datasets enables superior performance in various downstream tasks, such as question answering and sentiment analysis. BERT's bidirectional approach significantly advances contextual embeddings, marking a pivotal milestone in the evolution of language models.

4. Datasets

The training dataset provided during the track contains 9921 tweets extracted from [6] which contains anti-vaccine tweets from Twitter. It contains tweets along with the tweet IDs and the classes. The dataset represents a significant milestone as the inaugural large-scale compilation of approximately 10,000 COVID-19 anti-vaccine tweets, meticulously categorized into distinct anti-vaccine concerns within a multi-label framework. It stands out as the pioneering multi-label classification dataset, offering detailed explanations for each label. Notably, this dataset also includes class-wise summaries for all the tweets, providing a comprehensive resource for understanding and analyzing diverse anti-vaccine sentiments.

5. Preprocessing

To enhance the quality of the word embeddings generated by BERT, we conducted pre-processing on the tweets. Tweets inherently feature distinctive lexicons, such as hashtags, user mentions (@USER) and URLs (HTTP-URL). Without proper pre-processing, these elements can adversely impact the model's performance. Therefore, we implemented a meticulous data cleaning pipeline as an integral part of tweet pre-processing within our dataset.

1. Remove URLs: URLs do not help in multilabel classification; thus, we removed them with the help of regular expression from the text
2. Remove non-alphanumeric characters: We removed all the non-letter characters like brackets, colon, semi-colon, @, etc.
3. Remove Mentions: We removed all mentions as it might hinder the process of multilabel classification and is found often in tweets
4. HTML Tag Removal: We used BeautifulSoup to parse HTML and extract the text content, effectively removing any HTML tags.

5. Convert words to lower case: Tweets are written more casually, thus by lower casing every word, we are keeping only a single version of every word, enhancing the text analysis.

6. Methodology

6.1. COVID-Twitter-BERT (CT-BERT)

CT-BERT[5] is a specialized transformer-based model tailored to the domain of COVID-19 discourse. Pre-trained on an extensive dataset comprising tweets posted from January 12 to April 16, 2020, it initializes its weights using BERT-Large. Further refinement involves training on an additional 160 million tweets centered around the coronavirus topic. To ensure privacy, Twitter usernames were replaced with a standardized text token, and emoticons were substituted with English words. The selection of CT-BERT is strategic, aligning with the specificity of our training data, as opposed to BERT-Large trained on generic Wikipedia content, thereby enhancing the model's relevance and performance in the context of COVID-19-related tweet classification.

6.2. Model Summary

This model designed for multilabel classification, utilizing the CT-BERT[5] pre-trained model. Its architecture consists of a BERT layer, followed by a dropout layer (0.3), and concludes with a linear layer (12 outputs). The model extracts contextual embeddings from the BERT layer, applies dropout for regularization, and produces multilabel predictions through the linear layer. This design effectively adapts pre-trained knowledge to the specific multilabel classification task.

6.3. Tuning Parameter

The models have been run for 7 epochs with Adam optimizer[7] and the initial learning rate of $1e-5$. As no validation dataset was given, we divided the training data points into 80 and 20 split and used the 20 percent as a validation set. We predict the test set for the best validation performance.

7. Evaluation

The assessment of track results involves a meticulous evaluation, employing both the macro-F1 score and Jaccard index as a tie-breaker. In the context of the specified task, the outcome of our automated run is presented herein, reflecting a noteworthy achievement. Notably, my model clinched the top position, surpassing other submissions. The macro-F1 score, a key metric of performance, stands impressively at 0.71, while the Jaccard index, employed as a supplementary measure, registers at 0.72. This outcome underscores the effectiveness of our model, positioning it prominently within the competitive landscape of the task at hand.

Table 1
Results

Run File	Summary	Macro-F1	Jacc
final_df.csv	Fine-tuning Covid-Twitter Bert	0.71	0.72

8. Conclusion and future work

This study leverages the capabilities of Covid-Twitter-BERT, a transformer-based model pre-trained on an extensive corpus of COVID-19-related tweets. The primary objective is to conduct multilabel classification on anti-vaccine tweets, categorizing them into distinct domains encompassing conspiracy, country, ineffective, ingredients, mandatory, none, pharma, political, religious, rushed, side-effect, and unnecessary.

Our empirical findings underscore the superiority of transformer-based models, particularly Covid-Twitter-BERT, when contrasted with conventional natural language processing classifiers such as Naive Bayes, Logistic Regression, and Support Vector Machine[8]. The enhanced performance of transformer-based models emanates from their capacity to derive more expressive word embeddings, thereby yielding superior results across the designated classification task.

In pursuit of further refinement, we advocate for the exploration of data augmentation strategies to bolster the performance of our model. This strategic initiative is particularly pertinent given the inherent data-hungry nature of transformer-based models. Additionally, an avenue for potential improvement lies in the incorporation of adversarial training techniques, aimed at fortifying the model's robustness against diverse inputs and potential adversarial attacks. These proposed augmentations signify a commitment to continuous enhancement and resilience in the model's classification prowess.

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] L.-A. Cotfas, C. Delcea, I. Roxin, C. Ioanăș, D. S. Gherai, F. Tajariol, The longest month: Analyzing covid-19 vaccination opinions dynamics from tweets in the month following the first vaccine announcement, *IEEE Access* 9 (2021) 33203–33223. doi:10.1109/ACCESS.2021.3059821.
- [3] Z. Waseem, T. Davidson, D. Warmley, I. Weber, Understanding abuse: A typology of abusive language detection subtasks, in: Proceedings of the First Workshop on Abusive Language Online, Association for Computational Linguistics, Vancouver, BC, Canada, 2017, pp. 78–84. URL: <https://aclanthology.org/W17-3012>. doi:10.18653/v1/W17-3012.
- [4] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805* (2018).
- [5] M. Müller, M. Salathé, P. E. Kummervold, Covid-twitter-bert: A natural language processing

model to analyse covid-19 content on twitter, *Frontiers in Artificial Intelligence* 6 (2023) 1023281.

- [6] 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.
- [7] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, 2019. [arXiv:1711.05101](https://arxiv.org/abs/1711.05101).
- [8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2023. [arXiv:1706.03762](https://arxiv.org/abs/1706.03762).