VaxiBERT: A BERT-Based Classifier for Vaccine Tweets with Multi-Label Annotations

Shivangi Bithel, Samidha Verma, Prachi and Rajat Singh

Indian Institute of Technology, Delhi

Abstract

Vaccination has long been seen as an essential component of public health, providing a critical line of defense against infectious diseases. Our primary objective is to build a robust multi-label classification system capable of categorizing individual social media posts, specifically tweets, based on the numerous vaccine-related concerns stated by their writers. These reservations cover many issues, including misgivings about necessity, safety, and political intentions. We discuss our approach and evaluation results, shedding light on the intricate interplay between feeling, society, and science in the arena of the vaccine debate, using cutting-edge models such as Covid-Twitter-BERT and OpenLLaMA-7B. Our best-submitted run achieved a 0.67 macro-F1 Score and 0.70 Jaccard score.

Github Code: https://github.com/shivangibithel/VaxiBERT_AISoMe2023

Keywords

Sentiment Analysis, COVID-19 Vaccine Tweets, COVID-Twitter-BERT, Large Language Model, LoRA PEFT, Multi-label Classification

1. Introduction

A key component of public health for decades has been vaccination, a strong defense against the spreading of infectious diseases. Its significance in preventing outbreaks and safeguarding local populations cannot be emphasized. The crucial role that vaccination played in containing the COVID-19 pandemic, a global emergency that has brought vaccinations into the public eye like never before, has served as a reminder of the need for vaccination in the modern era. Beyond the pandemic reaction, widespread vaccination acceptance, especially on a societal level, continues to be essential in preventing disease resurgence, preventing childhood diseases, and reducing the yearly assault of seasonal illnesses like influenza.

However, the vaccine environment is defined by the complexity that reaches far beyond the scientific arena. A distinct undercurrent of suspicion has emerged, spurred by various issues ranging from politics to alleged side effects. This skepticism is a severe obstacle that must be addressed as we strive for widespread protection through vaccination. Understanding the



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

[☆] csy207657@cse.iitd.ac.in (S. Bithel); csy207575@cse.iitd.ac.in (S. Verma); prachi@cse.iitd.ac.in (Prachi); rajat.singh@cse.iitd.ac.in (R. Singh)

thtps://shivangibithel.github.io/ (S. Bithel); https://github.com/prach6i (Prachi); https://github.com/rajatb115 (R. Singh)

^{© 0000-0002-6152-4866 (}S. Bithel); 0009-0001-9513-3142 (S. Verma); 0009-0000-6663-5226 (Prachi); 0000-0002-9375-2580 (R. Singh)

^{© 0 2023} 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)

complex issues surrounding vaccinations is critical, and in this age of digital connectedness, social media platforms have emerged as a great source of information.

In this context, our work aims to navigate the complex web of public opinion on vaccines, mainly expressed in social media's unfiltered and dynamic arena. Our primary goal is to create a solid and dynamic multi-label classification system capable of categorizing individual social media posts, specifically tweets, based on the various vaccine-related concerns raised by their authors. It is critical to recognize that these worries are not uniform; a single tweet may include numerous separate vaccine-related concerns.

To simplify our classification task, we have a comprehensive set of concern labels that capture the wide range of anxieties permeating the vaccine discourse. These terms cover a variety of concerns, including skepticism about the necessity and safety of vaccinations, suspicions of larger conspiracies, political motivations behind vaccination mandates, and uncertainty about the effectiveness of vaccinations. Concerns also include vaccines' origins, make-up, and alleged negative effects, with personal religious beliefs influencing opinions.

In a time when information travels through digital channels at unprecedented speeds, our study aims to use the vast amounts of data generated on social media platforms to shed light on the complex world of vaccine apprehension. We aim to provide insightful contributions that can inform public health strategies, enhance vaccine communication, and foster a more nuanced understanding of the complex interplay between science, society, and sentiment in the field of vaccination by analyzing the concerns raised by individuals in their tweets.

2. Task

The task " Building 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" organized as a part of AISoMe (Artificial Intelligence on Social Media) Track in the FIRE (Forum for Information Retrieval Evaluation) 2023, we present an effective approach in this paper [1, 2]. A tweet can have more than one label (concern); e.g., a tweet expressing three different concerns towards vaccines will have three labels. The tweets are classified into multiple classes described below with examples:

- 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 the 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/politicians are pushing their own agenda through the vaccines."

- **Country** "Country of origin The tweet is against some vaccine because of the country where it was developed/manufactured"
- **Rushed** "Untested / Rushed Process The tweet expresses concerns that the vaccines have not been tested properly or that the published data is inaccurate."
- **Ingredients** "Vaccine Ingredients/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-effects** "Side Effects / Deaths The tweet expresses concerns about the side effects of the vaccines, including deaths caused."
- **Ineffective** "Vaccine is Ineffective The tweet expresses concerns that the 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."

Given below are a few examples of tweets along with their labels:

- "FYI....there are plenty of people walking around without vaccines for all sorts of contagious diseases/viruses. Why is Covid so different? We must ask why a mandatory vaccine card is even a consideration if the ones who are vaccinated feel that it protects them." : Mandatory Unnecessary
- "So there have been issues, but FDA are so desperate they deny it's the vaccine FDA reports facial paralysis in 4 volunteers for Pfizer's Covid-19 vaccine, but FDA denies vaccine is the cause Business Line https://t.co/nD8gwuxbvu" : **side-effects**
- "If this is seen as the deadliest disease in our lifetimes, and consequently the vaccine viewed as a miraculous panacea, why is Pfizer's stock price virtually unchanged from the beginning of the year?" : **pharma**
- "@MelanieMetz6 @XSOmegaMkII Inovio...look it up as well as Moderna. All 3 of these delivery methods have nano technology that can deliver DNA/RNA gene coding & mutation. This isnt a joke or up for speculation....its way beyond that now. I have leukaemia with 17q deletion So No." : side-effect ingredients conspiracy
- "Doctors Around the World Issue Dire WARNING: DO NOT GET THE COVID VACCINE!! https://t.co/JD5mlPTbVt via @Prepare_Change" : **none**
- "This is the same CEO that sold 60+% of his stock in Pfizer on the day of the vaccine announcement. Sell the news, don't take the vaccine, he seems super bullish on the long term successful prospects if this vaccine. https://t.co/m5dS8Y9Q9t" : **ineffective pharma**

3. Related Work

Users express their opinions regarding healthcare, diseases, treatments, vaccines, and immunization campaigns on microblogs like Twitter. In social computing, information extraction from these text-based tweets is increasingly popular. Classical machine learning techniques such as linear classifiers, Naive-Bayes classifiers, support vector machines, and deep neural techniques such as Long Short Term Memory(LSTMs) [3], Bidirectional RNN [4], BERT(Bidirectional Encoder Representations from Transformers) [5], and RoBERTa [6]. For natural language processing, more modern language models include large pre-trained models like T5 [7], GPT3 [8], LLaMA [9], PALM [10], and many more.

3.1. BERT

BERT [5] is a highly effective transformer-based architecture that adapts well to numerous tasks involving natural language processing. BERT allows for the pre-training of deep bidirectional representations from unlabeled text, which preserves more of the context and logical flow of the text. The model is pre-trained using next-sentence prediction (NSP) tasks and Masked Language Modelling (MLM). By including an additional output layer and achieving cutting-edge performance, the BERT model may be fine-tuned for a variety of jobs.

3.2. LLaMA

The Large Language Model Meta AI [9], abbreviated as LLaMA, represents a significant advancement in the realm of natural language processing. This collection of state-of-the-art foundation language models spans a spectrum of sizes, ranging from 7 billion to 65 billion parameters. What sets LLaMA apart is its ability to deliver exceptional performance while maintaining a comparatively smaller model size, thereby reducing the computational demands typically associated with cutting-edge language models. LLaMA's foundation models have been meticulously trained on a diverse and extensive range of unlabeled datasets. This training corpus includes data from sources such as CommonCrawl, C4, GitHub, Wikipedia, books, ArXiv, StackExchange, and more. The amalgamation of these varied datasets has empowered LLaMA to attain state-of-the-art performance, rivaling other top-performing models like Chinchilla-70B [11] and PaLM-540B [10].

4. Dataset

This work uses a training dataset created as part of the research project "CAVES: A dataset to facilitate explainable classification and summarization of concerns towards COVID-19 vaccines." [12]. This carefully managed training dataset includes a sizable corpus of 9,921 tweets criticizing the COVID-19 vaccination. These tweets were collected between 2020 and 2021 and have undergone meticulous manual annotation by subject-matter specialists. The issue categories in our research objectives have been carefully assigned to each tweet in this dataset. To assess the generalizability and robustness of our classification system, the test set encompasses approximately 500 tweets obtained from diverse sources. These tweets are not exclusively centered on COVID-19 vaccines; they span a broader spectrum, incorporating discussions on other vaccine types, such as the MMR and the flu.

5. Pre-processing

In line with prior research [13, 14], we conducted extensive pre-processing of the tweet data to enhance the quality of word embeddings. Tweets inherently feature unique lexicons, including

elements such as HASHTAGS, @USER mentions, HTTP-URLs, and EMOJIS. These elements often introduce noise if left unattended and adversely affect model performance. Therefore, we implemented a comprehensive data-cleaning pipeline as part of our tweet pre-processing procedure, encompassing the following key steps:

- **Stop Word Removal**: To streamline the text and emphasize essential information, we eliminated common stop words such as "the," "a," "an," and "in." These words typically do not contribute significant meaning to the text.
- **Lowercasing**: Given the informal nature of tweets, we converted all words to lowercase. This practice standardizes the text and ensures that each word is represented consistently, facilitating more effective text analysis.
- Emoticon Conversion: Emojis are frequently employed on Twitter to express emotions and sentiments. Recognizing their importance, we refrained from outright removal and instead converted emojis to their corresponding textual representations. This transformation retained the sentiment and emotional context of the text. The 'emoji' library (https://pypi.org/project/emoji/) aided in this process.
- **Contractions Expansion**: We systematically expanded contractions to their original, uncontracted forms to promote text standardization. For example, "don't" was expanded to "do not." We expanded this expansion by leveraging the 'contractions' library(https://pypi.org/project/contractions/).
- Non-Alphanumeric Character Removal: Extraneous non-letter characters, including brackets, colons, semi-colons, @ symbols, and the like, were removed from the text. This step contributed to text cleanliness and coherence.
- **URL Removal**: URLs unrelated to sentiment analysis were purged from the text using regular expressions. This exclusion aided in focusing the analysis on the textual content pertinent to sentiment assessment.

6. Methodology

• Run1: COVID-Twitter-BERT (CT-BERT): We used a domain-specific transformerbased model called CT-BERT[15]. We chose this model specifically because BERT-Large is trained on Wikipedia data, and using a pre-trained model in the same domain, in this case, COVID-19-related tweets would give more significant results after fine-tuning with the provided training data. We shuffled the training data, then split it into training and validation sets in the ratio 80:20 such that the percentage of instances of each class was preserved in both sets. Both training and validation instances were pre-processed, as explained in section 5. The resulting training data was used for fine-tuning CT-BERT[15] while validation data was used for evaluation. We trained the model for 15 epochs with a learning rate of 2e-5. The test data was also pre-processed using the same steps as training and validation data first to generate the embeddings for the tweet and then predict the probability scores of each tweet against all the classes. We used the sigmoid function over probability values with a threshold of 0.5 to predict the label. The final prediction file containing the Tweet ID and the predicted class was submitted as run1 for the task.

- **Run2: OpenLLaMA-7B**: We use the OpenLLaMA-7B [9] model variant to finetune for the task at hand. The training data instances were pre-processed, as explained in section 5. Following the same methodology as given in [16], we use the Prefix Tuning technique, which falls into the larger category of PEFT (Parameter Efficient Fine Tuning) approaches. In this, we learn a set of adaption prompt tokens, which is appended at the beginning of some top-N transformer layers. While finetuning, only these prompt tokens are finetuned for a specific downstream task, while the rest of the LLM parameters remain frozen. Also, a zero-initialized, zero-gated attention mechanism is used to inject the finetuned prompt token knowledge into the existing model so that the original model parameters don't deviate too much due to noise in the initial learning phase. We use 10 extra learnable prompt tokens in our setting and append them to the top 30 transformer layers. This adds only an extra 1.2M parameters over the existing 7B frozen parameters, requiring 10 minutes to train for 9,921 data points using batch size 4, 512 as max sequence length for 5 epochs using a learning rate of 9e-3. We generate the classification labels using the pre-processed test data as a text generation task.
- **Run3: OpenLLaMA-7B**: In run 3, we fine-tuned the OpenLLaMA-7B model variant for multi-label classification without pre-processing the tweets. All other details of model training are similar to the Run2. In this run, we generate the classification labels using the raw test data as a text generation task.

7. Evaluation

AISoMe Track results are evaluated using the macro-F1 score and Jaccard Score. The result of our three submitted runs for the task is shown in Table 1.

Sr No.	Team_ID	macro-F1 score	Jaccard Score	Rank
Run 1	DSIRC	0.67	0.7	4
Run 2	DSIRC	0.57	0.61	13
Run 3	DSIRC	0.55	0.6	16

Table 1

Result of AISoME Track Task

8. Conclusion and Future Work

This study employs Covid-Twitter-BERT and Open-LLaMA-7B to categorize vaccination-related tweets. The transformer-based model outperforms the fine-tuned OpenLLaMA-7B-based classifiers because its word embeddings are more expressive and yield better results on test data. Furthermore, because transformer-based models require many data, we recommend looking at data augmentation solutions to improve the performance of our model. Another aspect would be to train the model to become more robust against adversaries.

References

- 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.
- [2] 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.
- [3] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, J. Schmidhuber, Lstm: A search space odyssey, IEEE Transactions on Neural Networks and Learning Systems 28 (2015) 2222–2232. URL: https://api.semanticscholar.org/CorpusID:3356463.
- [4] D. E. Rumelhart, G. E. Hinton, R. J. Williams, Learning internal representations by error propagation, 1986. URL: https://api.semanticscholar.org/CorpusID:62245742.
- [5] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019. URL: https://aclanthology.org/N19-1423. doi:10. 18653/v1/N19-1423.
- [6] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019. arXiv:1907.11692.
- [7] C. Raffel, N. M. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, J. Mach. Learn. Res. 21 (2019) 140:1–140:67. URL: https://api.semanticscholar.org/CorpusID: 204838007.
- [8] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, 2020. arXiv:2005.14165.
- [9] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, ArXiv abs/2302.13971 (2023). URL: https: //api.semanticscholar.org/CorpusID:257219404.
- [10] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. M. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. C. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. García, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Díaz, O. Firat, M. Catasta, J. Wei, K. S. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, N. Fiedel, Palm: Scaling language modeling with pathways,

ArXiv abs/2204.02311 (2022). URL: https://api.semanticscholar.org/CorpusID:247951931.

- [11] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. van den Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, J. W. Rae, O. Vinyals, L. Sifre, Training compute-optimal large language models, ArXiv abs/2203.15556 (2022). URL: https://api.semanticscholar.org/CorpusID:247778764.
- [12] 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, SIGIR '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 3154–3164. URL: https://doi.org/10.1145/3477495.3531745. doi:10.1145/3477495.3531745.
- [13] S. Bithel, S. S. Malagi, Unsupervised identification of relevant prior cases, 2021. arXiv:2107.08973.
- [14] S. Bithel, Ctc: Covid-19 tweet classification using ct-bert (2022).
- [15] 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. arXiv:2005.07503.
- [16] R. Zhang, J. Han, C. Liu, P. Gao, A. Zhou, X. Hu, S. Yan, P. Lu, H. Li, Y. Qiao, Llama-adapter: Efficient fine-tuning of language models with zero-init attention, 2023. arXiv:2303.16199.