VaxVerdict :a RoBERTa-Based Multilabel Tweet Classifier

Sheetal Sonawane[1], Aditya Patil[1], Shivakumar Ranade[1] and Raj Awate[1]

[1]SCTR's Pune Institute of Computer Technology, Dhankawadi, Pune, 411043

Abstract

"VaxVerdict" is our submitted work to FIRE 2023 AISoMe Track Task. We propose using a multi-label classifier using roBERTa model to classify tweets as unnecessary, mandatory, pharma, conspiracy, political, country, rushed, ingredients, side-effects, ineffective, religious, and none. The rapid development of COVID-19 vaccines has marked a pivotal moment in the ongoing global effort to combat the pandemic caused by the SARS-CoV-2 virus. This has also spurred widespread discussions and debates on social platform like Twitter. The multilabel classification provides valuable insights into public opinion and their thinking towards vaccines. Monitoring and analyzing social media sentiments can help public authorities reform their communication and strategies to promote vaccines better. The evaluation score of our submitted predictions is reported in terms of accuracy and macro-F1 score. We achieved a Jaccard score of 0.67, a macro-F1 score of 0.65, and secured seventh rank among other submissions.

Keywords

COVID-19 Vaccine Tweets, Artificial Intelligence, Multi-label classifier, BERT, roBERTa

1. Introduction

The COVID-19 pandemic, an ongoing pandemic, has prompted the use of various medicines to control the spread of the virus and mitigate its symptoms. One of the most essential cures is the development and distribution of vaccines. As vaccination drives are being conducted, they derive a wide variety of public opinion ranging from enthusiasm to skepticism. Social media platforms like Twitter has served as a platform for people to share their thoughts and ideas regarding the vaccine. It allows public authorities to gain insights into the public's sentiments and attitudes toward vaccination. The insights gained from the analysis may help healthcare authorities, policymakers, and communicators frame better policies to regulate and foster vaccine confidence. The manual classification of tweets could be more varied and accurate. Hence, to classify the tweets, we will use a tweet dataset and classify them according to 12 classes- unnecessary, mandatory, pharma, conspiracy, political, country, rushed, ingredients, side-effects, ineffective, religious, and none. Then, we will use natural language processing techniques to analyze the sentiment expressed in the tweets. We will develop machine learning models that can help us classify tweets about the COVID-19 vaccines. By the end, we anticipate providing a detailed analysis of tweets, which will also uncover influential topics and regional nuances hidden within them.

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

sssonawane@pict.edu (S. Sonawane); adityapatilsy@gmail.com (A. Patil); shivakumar.ranade@gmail.com (S. Ranade); rajawate50@gmail.com (R. Awate)

2. Task

The goal is to build an effective multi-label classifier to label a social media post (mainly a tweet) according to the specific concern(s) towards vaccines, as expressed by the post's author. We consider the following concerns about vaccines as the labels for the classification task:

- 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 are against such companies 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/politicians are pushing their agenda through vaccines.
- Country: Country of origin — The tweet is against some immunization because of the country where it was developed/manufactured
- Rushed: Untested / Rushed Process — The tweet expresses concerns that the vaccines have not been tested correctly or that the published data is inaccurate.
- Ingredients: Vaccine Ingredients/technology — The tweet expresses concern about the ingredients present in the vaccines (e.g., fatal cells, chemicals) or the technology used (e.g., mRNA vaccines can change your DNA)
- Side-effect: Side Effects / Deaths — The tweet expresses concern about the side effects of the vaccines, including deaths caused.
- Ineffective: Vaccine is Ineffective — The tweet expresses concerns that the vaccines are useless and ineffective.
- 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

3. Related Work

Studies have found that sentiment varies over time and is influenced by factors such as s availability and vaccine efficacy. Others have found that most tweets express concerns about vaccines' safety and side effects. Tweets spreading false information about vaccines have a negative impact. Traditional machine learning methods like Naive-Bayes, Linear classifier, Support Vector Machine, Logistic Regression, Random Forest, and gradient boosting algorithms can be used for sentiment analysis. Deep neural techniques like Long Short-Term Memory (LSTMs), Bidirectional Recurrent Neural Networks (RNN), and Convolutional Neural Networks (CNNs) are very successful for text classification. More recent language models for natural language processing include BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer).

3.1. BERT

In natural language processing (NLP), BERT [2] (Bidirectional Encoder Representations from Transformers) is a potent and frequently used transformer-based approach. It has transformed NLP and is commonly used for various text-related tasks, including sentiment analysis. They are favored for challenging text analysis jobs like sentiment analysis of tweets about COVID-19 immunization because they are excellent at catching context and comprehending content. We can improve pre-trained BERT[2] models by using libraries like Hugging Face Transformers in Python, which offer simple access to pre-trained BERT[2] models and tools for fine-tuning.

4. Dataset

The training dataset provided during the track consists of 9921 tweets collected from various topics and areas over the internet. It contains the tweet IDs, the actual tweet content, the message, and the label given to that tweet. We used this dataset to train our model for predictions. Twelve unique labels in the dataset have been used to categorize the tweets based on their content.

4.1. Trends in Dataset

- Training dataset includes 14.7% side-effect, 12.68% rushed, 11.07% pharma,9.99% conspiracy, 9.85% mandatory, 9.72% unnecessary, 8.64% ingredients, 8.37% political, 8.23% ineffective, 4.32% country, 2.30% religious and 0.13% none tweets. All these represent the categories that are used to classify the tweets.
- Out of all the tweets, 80% of the tweets fall under a single category, and the remaining 20% fall under multiple categories.
- Amongst the tweets that fall under multiple categories, the most common categories are the rushed and side-effect categories. One hundred twenty tweets are classified as belonging to both these categories.
- 9191 out of 9921 tweets contain the word "vaccine" in them, indicating the data quality the model is trained on. Only the remaining 730 tweets do not include "vaccine" or any of its forms and do not directly refer to vaccination.
- Out of the 9921 tweets, 8540 mentioned COVID-19, and the remaining 1381 did not.
- 56.83% of the tweets mention the names of big vaccine manufacturers such as Pfizer, AstraZeneca, Johnson & Johnson, Moderna, and a few others.

5. Preprocessing

We pre-processed the tweets to improve the quality of word embeddings produced by BERT [2]. Tweets generally contain many unwanted symbols that need to be removed to generate better word embeddings.

- Removing punctuation used in the tweets: The tweets contain lots of unnecessary punctuation marks that add no significant meaning to the sentence and need to be removed. It

involves the removal of punctuations normally used in tweets, such as commas(,), exclamation marks (!), apostrophes (’), double quotes("), question marks(?), etc. It improves the overall consistency of the text. It reduces the text’s noise, making it easier to focus on the essential content and semantics. Also, removing punctuation assists in tokenization because it separates words from punctuation marks, leading to more accurate tokenization. It also makes stemming or lemmatization of the text easier and improves parsing of the text.

- Removing symbols: Tweets contain many characters like ’@’ or ’$’ or ’%,’ or ’’ etc., which don’t contribute much to the meaning of the text and are just used to represent certain entities. Removing symbols reduces noise and improves overall consistency. It also leads to improved analysis; if not removed, additional tokens unrelated to the text will be introduced, potentially producing less meaningful results.
- Conversion of text to lowercase: This is an essential preprocessing step that we performed to bring uniformity and consistency to the reader. Expressing a word in lowercase, uppercase, or camelcase does not change its meaning, but the model will learn the term differently and might affect the outcome. So, to remove this ambiguity, converting the text to lowercase is a crucial step. It also leads to reduced vocabulary size as the words "apple" and "Apple," although having the same meaning, will be treated as separate words, increasing the vocabulary size and increasing redundancy. Conversion to lowercase also improves text matching and improves overall efficiency.

6. Methodology

6.1. Model

RoBERTa [1], short for "A Robustly Optimized BERT Pretraining Approach," is a state-of-the-art natural language processing (NLP) model. It builds upon the foundation of BERT (Bidirectional Encoder Representations from Transformers) and is designed for various NLP tasks, including text classification, sentiment analysis, and language understanding. RoBERTa[1] is known for its robustness and effectiveness in understanding context and semantics within text data. It achieves this by pretraining on a massive amount of publicly available text from the internet, effectively learning to comprehend diverse and complex language patterns. One key feature of RoBERTa [1]is its training methodology, which includes large-scale data, longer training times, and extensive hyperparameter tuning. These factors contribute to its superior performance on various NLP benchmarks and real-world applications. In our research, we utilized the RoBERTa [1]model as the foundation for our tweet classification task, leveraging its ability to capture nuanced meanings in text data, particularly tweets related to COVID-19 vaccines. The model’s contextual understanding and fine-tuned architecture played a pivotal role in achieving accurate and effective tweet classification.

6.2. Experimental Setup

- Data Split: We divided the dataset into three sets: a training set, a validation set, and a test set. The training set comprised 70% of the data, while the validation and test set

accounted for 15% . This split was performed to ensure appropriate model training and evaluation. We used the train_test_split function from scikit-learn, and the random seed was set to 42 for reproducibility.

- Training Loop: Model training was carried out throughout 20 epochs. We adopted the AdamW optimizer and the Binary Cross-Entropy With Logits Loss (BCEWithLogitsLoss) as the loss function. The training loop involved iterating through batches of data, computing gradients, and optimizing the model's weights. The training aimed to minimize the loss function and enhance the model's classification performance.

6.3. Prediction

We employed the trained RoBERTa[1] model as a Multilabel Classifier for the prediction phase to classify the test tweets into one or more of the 12 target classes. The process can be summarized as follows:

- Generating Embeddings: We utilized the RoBERTa[1] model to generate embeddings for the test tweets. These embeddings capture the underlying semantic information in the text, enabling the model to understand the context of each tweet effectively.
- Sigmoid Values: The output layer of the RoBERTa[1] model consists of neurons corresponding to the 12 target classes ('unnecessary,' 'mandatory,' 'pharma,' 'conspiracy,' 'political,' 'country,' 'rushed,' 'ingredients,' 'side-effect,' 'ineffective,' 'religious,' and 'none'). We predicted sigmoid values for each output neuron, representing the likelihood of a tweet belonging to a particular class. These values ranged between 0 and 1, indicating the model's confidence in each classification.
- Thresholding: To make the final class predictions, we applied a threshold of 0.5 to the sigmoid values. If a sigmoid value for a particular class exceeded 0.5, the tweet was considered to belong to that class; otherwise, it was not assigned to that class.
- Submission: The outcome of this process was a set of predicted classes for each test tweet. We compiled these predictions into a final prediction file, including the Tweet ID and the corresponding predicted classes. This prediction file was submitted as our run for the Task evaluation.

7. Evaluation

The task for AISoMe Track results is evaluated using the macro-F1 score on the twelve classes as metrics. The result of our submitted automated run for Task is shown in Table 1. [Heading] got the 7th rank among other submissions with the macro-F1 score of 0.65. The ranks and the macro F1 scores of the other two run submissions with their respective models used are also shown below.

Table 1
Results of Task

Sr. No.	Run File	Model	Macro-F1	Jacc	Rank
1	predictions (2).csv	RoBERTa	0.65	0.67	7
2	predictions (1).csv	BERT	0.6	0.65	12
3	predictions.csv	CNN	0.55	0.6	17

8. Conclusion and Future Work

In our quest to advance "VaxVerdict" and contribute to the ongoing progress of multilabel tweet classification, several promising avenues for future work emerge:

- Hyperparameter Tuning: To further optimize the performance of our model, we recommend an in-depth exploration of hyperparameter tuning. Fine-tuning hyperparameters such as learning rates, batch sizes, and optimizer configurations can potentially yield significant improvements in model accuracy and convergence speed.
- Data Augmentation: Given the data-intensive nature of transformer-based models, research into data augmentation strategies tailored specifically for social media text remains an essential area of exploration. Augmentation techniques can expand the training dataset, potentially enhancing the model's generalization.
- Adversarial Training: Continuing the investigation into adversarial training methods can fortify the model's robustness, ensuring it can handle noisy and adversarial inputs present in social media content.
- Checkpoint Incorporation: Implementing model checkpoints at strategic intervals during training can provide resilience against training interruptions and hardware failures. This ensures that valuable training progress is preserved and can be resumed efficiently.

References

[1] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov. *RoBERTa: A Robustly Optimized BERT Pretraining Approach.* ArXiv, 2019. https://arxiv.org/abs/1907.11692

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* ArXiv, 2018. https://arxiv.org/abs/1810.04805

[3] Martin Müller, Marcel Salathé, Per Egil Kummervold. *COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter.* ArXiv, 2020. https://arxiv.org/abs/2005.07503

[4] Soham Poddar, Moumita Basu, Kripabandhu Ghosh, Saptarshi Ghosh. *Overview of the FIRE 2023 Track: Artificial Intelligence on Social Media (AISoMe).* Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation, 2023.

[5] Soham Poddar, Azlaan Mustafa Samad, Rajdeep Mukherjee, Niloy Ganguly, Saptarshi Ghosh. *CAVES: A Dataset to facilitate Explainable Classification and Summarization of*

Concerns towards COVID Vaccines. Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 3154–3164.