VaccineBERT: BERT for COVID-19 Vaccine Tweet Classification

Shivangi Bithel[1], Samidha Verma[2]

[1]Indian Institute of Technology, Delhi, Hauz Khas, New Delhi, Delhi 110016
[2]Indian Institute of Technology, Delhi, Hauz Khas, New Delhi, Delhi 110016

Abstract

VaccineBERT is our submitted work to FIRE 2021 IRMiDis Track Task 2. We propose using a domain-specific BERT model to classify tweets as ProVax, AntiVax, and Neutral. The vaccination process is ongoing worldwide to fight against the novel coronavirus disease (COVID-19), and the sentiment analysis of tweets can provide helpful insights regarding the stance of people about the new vaccine. Governments can plan their strategies based on people's points of view about the vaccine to make the vaccination drives successful. The evaluation score of our submitted run is reported in terms of accuracy and macro-F1 score. We achieved an accuracy of 0.576, a macro-F1 score of 0.582, and enjoyed the first rank among other submissions.

Keywords

Sentiment Analysis, COVID-19 Vaccine Tweets, COVID-Twitter-BERT

1. Introduction

Today the world is fighting its most challenging battle in the form of the COVID-19 pandemic. Over the years, vaccines have been proven to be a very safe and effective way to fight and eradicate infectious diseases by providing immunity to people to fight against viruses. Thus a race to discover new and effective vaccines made it possible to provide the Corona virus vaccine to the world so soon. People are using social media sites like Twitter to discuss about the vaccine as it is being distributed around the globe. The discussions of vaccination progress, accessibility, efficacy, and side effects are ongoing, and people have both positive and negative views about it. It is helpful for the government and various health organizations like WHO to know what people think about the new COVID-19 vaccines. They can use the insights from these micro-blogs to plan their future strategies and encourage everyone to get fully vaccinated.

It is complex but also imperative to stop the spread of misinformation about the COVID-19 vaccine. The government is trying to stop the pandemic as well as the growing infodemic around the vaccine from spreading. Twitter also tries to ban tweets that involve incorrect or misleading information about the virus, its preventive measures, and treatments. Manual classification of tweets is tedious and erroneous. Hence, there is a desperate need to develop the machine learning models that can help us in the task of classifying tweets about the COVID-19 vaccines.

Forum for Information Retrieval Evaluation, December 13-17, 2021, India
csy207657@cse.iitd.ac.in (S. Bithel); csy207575@cse.iitd.ac.in (S. Verma)
https://shivangibithel.github.io/ (S. Bithel); https:// (S. Verma)
0000-0002-6152-4866 (S. Bithel); 0000-0000-0000-0000 (S. Verma)

2. Task

For task 2, "Building an effective classifier for 3-class classification on tweets regarding people's stance towards COVID-19 vaccines", organized as a part of IRMiDis (Information Retrieval from Microblogs during Disasters) Track in the FIRE (Forum for Information Retrieval Evaluation) 2021, we present an effective approach in this paper. The tweets are classified into 3 classes described below with examples:

- AntiVax - "The tweet is against the use of vaccines."
- ProVax - "The tweet supports / promotes the use of vaccines."
- Neutral - "The tweet does not have any discernible sentiment expressed towards vaccines or is not related to vaccines."

An example for each class of tweets has been given below:

- **AntiVax Tweet:** *"They can have their vaccine, I want the right to say no- not in my body. We will only have that right under Donald Ĵ Trump https://t.co/MrfDSMm6ĴB "*
- **ProVax Tweet:** *"Best news of the year so far, well at least for the last 34 weeks! One of the many vaccines against COVID19 being developed and looking extremely positive. We can start to see the light for 2021!!"*
- **Neutral Tweet:** *"Will you REFUSE the Pfizer vaccine even if it means losing your job?"*

3. Related Work

Users post content on microblogs like twitter for various purposes, including their sentiments about Coronavirus, COVID-19 vaccines and vaccination drives. Information extraction from these textual tweets is very popular part of social computing. 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 includes **BERT** (Bidirectional Encoder Representations from Transformers) [1] and its domain-specific version **CT-BERT** (COVID-Twitter-BERT) [2].

3.1. BERT

BERT is a very powerful transformer-based architecture that generalizes well to many natural language processing tasks. Using BERT, deep bidirectional representations can be pre-trained from unlabeled text, which retains more information about the context and flow of the text. Model is pre-trained using Masked Language Modelling (MLM) task and Next Sentence prediction task. The BERT model can be fine-tuned for various tasks by adding an additional output layer and giving a state-of-the-art performance..

4. Dataset

The training dataset provided during the track contains 2792 tweets extracted from [3] on the stance of people towards COVID-19 vaccine crawled between November-December 2020. It contains tweets along with the tweet IDs and the classes. The test dataset contains 1600 tweets crawled using vaccine-related terms between March-December 2020. It contains tweets along with tweet IDs. Our approach used the dataset by Muller et al. [4] and crawled Twitter for more information. We augmented the dataset by Muller et al. [4] with attributes like screen name, retweet count, followers count, friends count, status count, verified status, and name of the user associated with the given tweet and tweet ID by using python API Tweepy [5] to observe the various trends in the data.

4.1. Trends in the dataset

Based on the given and collected information, the following trends were observed in the dataset.

- Training dataset includes 36.1% Neutral, 35.5% ProVax, and 28.4% AntiVax tweets.
- People with more than 10000 followers tend to post 58.1% Neutral, 32.7% ProVax, and 9.2% AntiVax (very less) tweets.
- People having verified accounts on Twitter tend to post 49.4% Neutral, 41.0% ProVax, and 9.5% AntiVax (very less) tweets.
- People with more than 1000 friends tend to post 37.2% Neutral, 36% ProVax, and 26.8% AntiVax tweets.
- People with more than 10000 status tend to post 42.6% Neutral, 31% ProVax, and 26.4% AntiVax tweets on their wall.
- Most common words in the dataset include "vaccine", "vaccines", "covid19", "news", "Pfizer" and "coronavirus" having more than 500 mentions.
- Most common accounts tagged in the tweets include "@realdonaldtrump" and "@pfizer", with more than 50 mentions.

The test data is annotated by three human annotators, where a label is assigned on the unanimous agreement or majority agreement (2 out of 3) from the given labels.

5. Pre-processing

Following the prior experience of NLP tasks[6], we pre-processed the tweets in order to improve the quality of word embeddings produced by BERT. Tweets generally contain unique lexicons like *HASHTAGS*, *@USER*, *HTTP-URL* and *EMOJIS* which without pre-processing, often reduce the performance of the model. Thus, we used the following data cleaning pipeline as part of pre-processing the tweets in the dataset:

- **Remove stop words:** A stop word is a commonly used word such as "the", "a", "an", "in", which do not provide any valuable information. We remove the stop words in order to give more focus to the important information.

- **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.
- **Convert emoticons to words:** Emojis are extensively used on Twitter to express feelings and emotions. Completely removing them removes a lot of sentiment information; thus, we converted the emojis to text and retained their meaning using 'emoji' library (https://pypi.org/project/emoji/).
- **Expand contractions to text:** In order to standardize our text, each contraction is converted to its expanded, original form. We used the 'contractions' library (https://pypi.org/project/contractions/) to expand the words like "don't" to "do not".
- **Remove non-alphanumeric characters:** We removed all the non-letter characters like brackets, colon, semi-colon, @, etc.
- **Remove URLs:** URLs do not help in sentiment analysis; thus, we removed them with the help of regular expression from the text.

6. Methodology

6.1. Model

COVID-Twitter-BERT (CT-BERT): CT-BERT[2] is a domain-specific transformer-based model, pre-trained on a large corpus of tweets posted between January 12 to April 16, 2020, on the topic of COVID-19. It uses the BERT-Large weights for initialization and further pre-trained over 160M tweets about the coronavirus. The tweets were pseudonymized by replacing all Twitter usernames with a common text token. English words also replaced all the emoticons in the tweets. We specifically used this model since BERT-Large is trained on Wikipedia data, and using a model that is pre-trained in the same domain, i.e., Covid-19 related tweets, in this case, would intuitively give better results upon fine-tuning with the given training data.

6.2. Experimental Setup

We first 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 were 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[2] while validation data was used for evaluation. In order to prevent overfitting, we used early stop monitoring the validation loss with patience value 3.

6.3. Prediction

For the prediction over the available test data, we used the fine-tuned CT-BERT model as a text classification model to generate the embeddings for the tweet and then further predict the probability scores of each tweet against all three classes. The class having the maximum probability was reported as the predicted class for that tweet. The final prediction file containing the Tweet ID and the predicted class was submitted as run for Task 2.

7. Evaluation

Task 2 - IRMiDis Track results are evaluated using overall accuracy and the macro-F1 score on the three classes as metrics. The result of our submitted automated run for Task 2 is shown in Table 1. VaccineBERT got the 1st rank among other submissions with an overall accuracy of 0.576 and the macro-F1 score of 0.582.

Sr No.	Team_ID	Accuracy	macro-F1 score	Rank
1	IR_IITD	0.576	0.582	1

Table 1
Result of Task 2

8. Conclusion and Future Work

This paper uses Covid-Twitter-BERT, a transformer-based model pre-trained on a large corpus of COVID-19 related tweets, to classify tweets as ProVax, AntiVax, or Neutral. We observed that the transformer-based model outperformed the traditional natural language processing classifier, namely Naive Bayes, Logistic Regression, and Support Vector Machine, as word embeddings computed by the former are more expressive and yield better results on the task. We further propose to look into data augmentation strategies for improving the performance of our model since transformer-based models are data-hungry. Another addition could be to adversarially train the model to improve its robustness.

References

[1] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. URL: http://arxiv.org/abs/1810.04805.

[2] M. Müller, M. Salathé, P. E. Kummervold, Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter, arXiv preprint arXiv:2005.07503 (2020).

[3] L.-A. Cotfas, C. Delcea, I. Roxin, C. Ioanăş, 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.

[4] M. M. Müller, M. Salathé, Crowdbreaks: Tracking health trends using public social media data and crowdsourcing, Frontiers in Public Health 7 (2019) 81. URL: https://www.frontiersin.org/article/10.3389/fpubh.2019.00081. doi:10.3389/fpubh.2019.00081.

[5] J. Roesslein, Tweepy: Twitter for python!, URL: https://github.com/tweepy/tweepy (2020).

[6] S. Bithel, S. S. Malagi, Unsupervised identification of relevant prior cases, 2021. arXiv:2107.08973.