Natural Language Processing to Analyze COVID-19 Anti-vaccination Tweets

Jeongyeob Hong

COMP 484: Introduction to Artificial Intelligence

Prof. Susan Fox

Fall 2021

Abstract

Social media has provided a global medium for the spread of misinformation. Our research pertains to the misinformation circulating Twitter regarding the COVID-19 vaccine, aiming to test the performance of natural language processing to conduct fine-grained analysis of Twitter data regarding COVID-19 vaccines. Our research aimed to test the question of whether or not a BERT (Bidirectional Encoder Representations from Transformers) model can classify intention behind COVID-19 vaccine tweets based on text content. A distilled BERT model trained on 1.2K human-annotated tweets into categories of content regarding anti-vaccination sentiments (government conspiracy, threats to medical freedom, and medical risks of vaccination) ran with an average 75% accuracy in identifying anti-vaccination sentiments in Twitter data.

Introduction and Project Plan

Since the approval of the COVID-19 vaccine by the FDA, there has been a considerable amount of animosity and distrust towards the vaccine by a group that is known in popular culture as “anti-vaxxers.” Anti-vaxxers take to social media sites, such as Twitter and Facebook, and spread misinformation about the vaccine, which has become a politically polarizing subject in today’s culture. When misinformation sows seeds of doubt into peoples’ minds, vaccination rates are less likely to be able to fight against COVID-19. Our project seeks to create an intent detection program that will detect the intentions behind Tweets spreading misinformation about the COVID-19 vaccine. We find this work to be important because intent analysis on misinformation could reveal what kind of information is spread on the internet through ill-intent and for what purposes this misinformation is spread. Our group broke apart the different categories of misinformation specifically about the COVID-19 vaccine to analyze a dataset of tweets from 2020-2021. Our goal for this project is to be able to create a model that can at least return an intention. We will not create or restructure a new model. Instead, we will use fine tuning for pretrained models. The dataset that we annotated will be used for a fine tuning process. The success of this model depends on its accuracy calculated by comparing a validation set.

Our dataset consists of more than million tweets scrolled based on keywords or hashtags. From 56 key words, we came up with multiple categories: personal story, risk of vaccine, conspiracy theory, and medical freedom, as general types of tweets found in the dataset that gave misinformation about the vaccine. Among millions of tweets, we will first start with 10K datasets. 2.5K will be test data and 7.5k will be the training dataset. A validation dataset will either be added or use some portion of the training dataset. All of the annotation will be done by team members, based on an annotation guideline, which is written below. After pre-processing the dataset, we will use a transformer-based model. Starting with the existing transformer model (BERT), we will eventually fine-tune the model to better suit our purpose. Since the initial dataset is not extremely large, we will run the model in Google Colab using cpus. Most of the team members are familiar with Tensorflow; therefore, we will use Tensorflow and supporting libraries such as Numpy, Keras, and etc. We will write our own codes for the fine-tuning process as well as post-processing for demonstration.

Background Knowledge

One of the major areas of research and commercial use of natural language processing today is sentiment analysis. Sentiment analysis is a process of text classification that categorizes texts based on the sentiments that are expressed in a text. When applied with machine learning, sentiment analysis can be used to make predictions about text sentiments based on training data. Typically, sentiment analysis is limited to analyzing text as either positive, negative or neutral. A more specialized variation of sentiment analysis is intention analysis, which goes a step beyond just positive, negative, or neutral categorization and instead identifies the intention within a text. Intention analysis opens up the capabilities of natural language processing to be able to understand speech discourse, but it also provides an opportunity for companies to use online consumer data to understand how consumers interact/engage with a product or company. Categories for intention analysis in this context include intents to: *purchase*, *inquire*, *complain*, *criticize*, *praise*, *direct*, *quit*, *compare*, *wish* and *sell*. (Carlos) Intention analysis was the basis that we used to create our fine-grained classification model for our dataset of COVID-19 vaccine tweets. Annotating the dataset based on the categories that were chosen to classify the kinds of anti-vaccination tweets would allow the model to analyze the text and capture the overall messaging that the tweet implies.

Relevant work

The inspiration for this project came from a proposed study by Chen et al. on using intent detection on a corpus of tweets about the COVID-19 pandemic, with a focus on using their machine learning model as a way to identify false information. The methodology of this study involved classifying intent within three categories of false information: disinformation, misinformation, and malinformation. Disinformation was defined as “information that is false and deliberately created, where the intent behind it might be harmful to individuals or groups”; misinformation as “information that is false, but the intent behind it is not harmful”; and malinformation as “ information that is sometimes based on reality, but is intended for causing harm.” Within these categories Chen et al placed several examples of applications of these types of false information. This study was published during the earlier stages of the COVID-19 pandemic and the dataset studied focused on content related to the pandemic itself. Our team wanted to explore tweets related to the COVID-19 vaccine, since this topic was understudied and is relevant to discussions going on currently about public health, which would provide more varied intentions against vaccines than current sentiments regarding the pandemic as a whole.

The methodology for our project was inspired by another study which investigated the performance of natural language processing to learn classification of emotion in text. In a study by Demszky et al, the researchers used a BERT model and transfer learning to train a model to identify 27 different emotion categories in a large text dataset of 58 thousand human-annotated Reddit comments. The results of this study concluded that though the BERT model performs well with fine-grained categorization, their study left much room for improvement. Our group decided to use the BERT model with more general categories for classification, as this paper showed that fine-tuning BERT led to less accurate results due to the constant need to generalize emotional categories. Nevertheless, the study provided a baseline for annotation guidelines and dataset validation.

The dataset used for training and testing was developed by Muric et. al, to “better understand the phenomenon of vaccine hesitancy through the lens of social media.” The dataset contains about 38 million tweets that were collected between 2020 and 2021, although our project opted to only use tweets from months following the approval of COVID-19 vaccinations in 2021, in order to have enough tweets that would show the most recent developments of reasoning behind anti-vaccination sentiments. Due to the thoroughness of the dataset, it leaves open the possibility of tracking the sentiments towards the vaccine and the kinds of arguments or reasonings that some may be against vaccination.

Finally, the annotation guidelines for the dataset previously mentioned were constructed based on the information provided by Sanchez et al. in a study titled “Analysis of the Anti-Vaccine Movement in Social Networks.” This paper provided an analysis on the ways in which anti-vaccination misinformation gets spread on different social media platforms. Although the material studied in this study is not specific to COVID, a lot of the mechanisms for spreading misinformation about vaccines by people online are reflected in the anti-vaccine movement seen today against the COVID vaccine. Namely, vaccine risks, personal stories, spreading of conspiracy theories, and mistrust of the pharmaceutical industry, were highlighted in this paper, and would go on to be used as categories for our project’s intent classification.

Our Approach

In this section, we will go over our codes, explain our reasoning for the decisions we made, and justify them. Setup, preprocessing, tokenization, compile and train, and demo are the five sections of our code. During the setup stage, we installed the required libraries in Google Colab and linked our results to Google Drive. We were able to retrieve data more efficiently as a result of this because of its compatibility with the pickle data form. This is also the reason we used the Pandas library to load and manage data. The goal of the entire process was to reduce data management time and costs.

We create three functions in the preprocessing section. The `convertAscii` function is used to convert Unicode to Ascill. We used this in the hopes of deleting emoji and removing cra data. The `cleanStopAndShortWords` function removes two letter words and two English stopwords. Stop words and short words are commonly deleted in NLP because they do not provide highly valuable information. To clean the data, the `preprocess` function employs the previous two functions as well as regex expressions. We had to remove any user names due to the nature of our dataset as well as emojis and URL links.

The next step is to tokenize the data and prepare it for the Bert model. We will go over this model in detail in the following section, but for now, keep in mind that we covered a mask on some of our input sentences. We now create a pickle file in our drive after printing some data facts so that we can access the model data whenever we want. The model is then compiled and trained. We tested the performance of our model using Adam optimizer and testing many learning rates with different batch sizes. The `Earlystop` function is an intriguing feature of TensorFlow that we used. If the validation loss stops decreasing after a certain number of epochs, the training is stopped. The demo file retrieves the trained model on the shared drive. Tensorflow's save and load functions were not used. Instead, we only brought weights and transferred to a new model. We have a function called `sentimental_analysis` in the demo file that takes input text and a trained model. The model predicts the input sentence. This task's output is the tensor itself. So we used an Encoder to convert the numeric tensor to natural language.

We decided to use the Bert model because of its bidirectional feature. Prior to Bert, NLP was dominated by sequential models such as Recurrent Neural Network (RNN). The main issue of RNN and sequential networks is loss of features from embedding vectors as sentences get bigger. Therefore, the output statements were often inconsistent with the subject, which positions at the beginning of the sentence. The model's failure of incorporating contexts was the main issue of the NLP industry.

To solve this issue, researchers created bidirectional LSTM basically using two models which read the input sentence in opposite directions: one goes from left to right and the other goes the opposite. However, this approach turned out to be limited. ELMO, the state of the art model prior to BERT, was based on bidirectional LSTM. Its rival, GPT2, was using a Transformer. Transformer is a sequential model with a special function called attention. After each embedding vector is forwarded into the hidden state, they apply an attention function to calculate the weight, basically asking the model to pay 'attention' to certain stages of deep learning layers. Transformer improved the performance of models but it was not enough. GPT2 utilized only the decoder of transformers and used numerous hyperparameters. However, BERT broke all of the SOTA records by using two pre-trained tasks: Masking and Next sentence prediction.

Masking the input refers to an act in which a model makes a random token of input sentence blank. For example, so far models have received input sentences like "I love this but I am not going to major in Computer Science". BERT received a previous sentence as well as "I **MSK MSK** but I am not going to major in **MSK** Science". By comparing two inputs, the model learns which surrounding words need more attention. This was BERT's way of solving contexts inside the text.

Then BERT tried to master contexts or transitions between sentences. Next sentence prediction task comes in here. The model receives input of two sentences and predicts whether the second sentence is an adequate follow up of the previous one. Half of the training examples were wrong, so the model learned a pattern from a well-distributed dataset.

By only using the encoder of the transformer with two pre-trained tasks that aim to master contexts, BERT was able to achieve the SOTA results. For our project, understanding context was crucial. Tweets are often referring to each other and a lot of them have sarcasms as well. That is why we used the best model that captures the context. Due to overfitting issues, we had to choose DistillBert, which is basically the same BERT model with less hyperparameter, layers and training datasets.

When it comes to annotating data, the most important factor is IAA (Inter-Annotator Agreement). As its name suggests, IAA refers to the consistency between annotators and its metric, IAA score, is crucial for evaluating the dataset. Good IAA score starts from concrete guidelines. Therefore, we came up with a guideline for this project. Out of all approaches, we decided to focus on keywords, since it was the easiest given that our dataset provides hashtags. However, we do acknowledge that this will make the model act as if it is rule based.

Analysis

We made several modifications and improvements to our program during our testing phase for the project. Through each iteration of testing we also added more annotations to our dataset, which also affected the accuracy of our model. In this section, we will discuss each of the improvements made to the model and the results of those improvements, as well as the factors that we felt influenced the effectiveness of our model and how well it performed with our own test inputs.

The first time that we trained the BERT model, we encountered several areas for improvement. The accuracy of the model was below 50%, which we mostly attributed to the insufficient data that we had when we first trained the model. As shown in the results below, the deviation in growth between the accuracy and validation accuracy of the model began to show overfitting. At this time, we attributed these results mostly to the lack of annotated data. Of the 1000 final annotated data that we had when we began training, only 433 were usable for training. The learning rate for this model was set to 2e-5, which resulted in lower accuracy and high loss. To remedy this, the learning rate was set to 1e-4 and the model was retrained using the same dataset. The results were more favorable, however the model began to show overfitting after the second epoch in training. These first results suggested that more data was needed to improve accuracy and avoid overfitting.

	accuracy	val_accuracy	loss	val_loss
Epoch 1	0.6171	0.6705	0.9734	0.9139
Epoch 2	0.7343	0.6477	0.7703	0.8625
Epoch 3	0.8086	0.6136	0.6215	1.0536
Epoch 4	0.8486	0.6477	0.4873	1.0311

Figure 1. Learning rate 1e-4

	accuracy	val_accuracy	loss	val_loss
Epoch 1	0.4171	0.3636	1.2695	1.2108
Epoch 2	0.5629	0.5682	1.1434	0.6477
Epoch 3	0.6571	0.6354	1.0240	1.0086
Epoch 4	0.7000	0.6477	0.9168	0.9242

Figure 2. Learning rate 2e-5

Annotating more data points from our set and using a smaller and simpler model proved to be useful improvements upon the second round of testing. The DistilBERT model has the same model architecture as the BERT model, but utilizes less parameters and is a smaller model than the general BERT model. According to a study by Sanh, the DistilBERT model is able to reach similar performance levels compared to the larger BERT model, while being able to train on less data. To test the impact that this would have on our project, we trained and tested a DistilBERT model using the same data that was used in the previous round of training and testing. The results from this are shown in the figure below. As expected, the DistilBERT model resulted in less overfitting throughout the four epochs and similar levels of accuracy overall compared to the larger BERT model. In the end DistilBERT is the preferred model for the purposes of this project.

	accuracy	val_accuracy	loss	val_loss
Epoch 1	0.3914	0.4318	1.3113	1.2004
Epoch 2	0.4886	0.6477	1.1762	1.0899
Epoch 3	0.6314	0.6591	1.0539	0.9508
Epoch 4	0.6657	0.6250	0.9004	0.8407

Figure 3. Learning rate 2e-5

In order to test the model built with DistilBERT, the model was trained with a second set of readily annotated text data and the accuracy and loss values from the training epochs were compared to the results that were obtained from the BERT and DistilBERT trials. The results from that test are shown in Figure 4 below:

	accuracy	val_accuracy	loss	val_loss
Epoch 1	0.6567	0.7467	0.7667	0.6167
Epoch 2	0.7201	0.6384	0.6737	0.8446
Epoch 3	0.6991	0.7509	0.6668	0.6646

Figure 4.

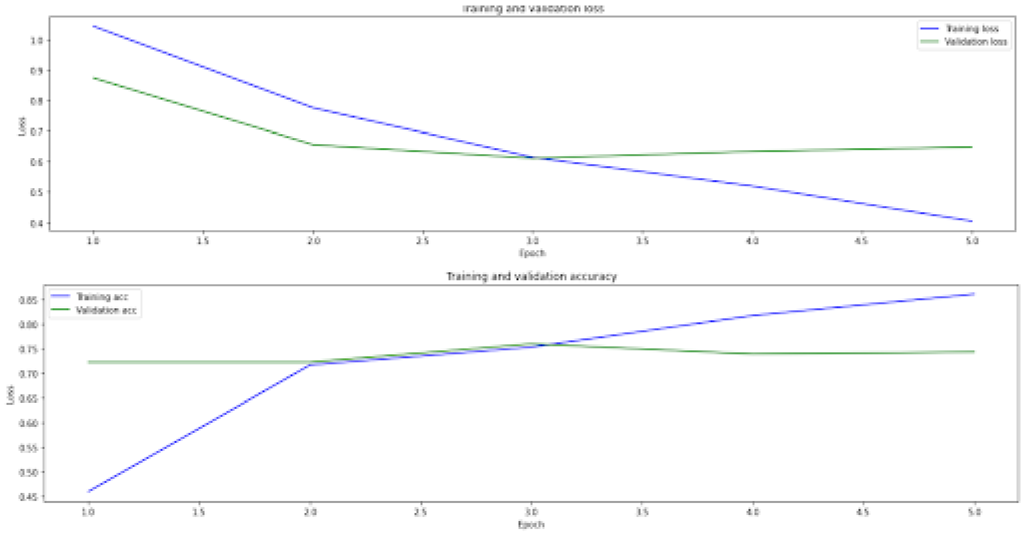
An additional adjustment made to improve the model's accuracy involved removing one of the categories of data from our training set. The "personal anecdote" category, which meant to categorize tweets that did not necessarily have an intention of spreading misinformation but still had negative connotations related to COVID vaccination, did not have enough tweets from the data that our team annotated to have any significance during training. In addition, many tweets that were told as anecdotal experiences with vaccination ended up falling into other intention categories, and were more prevalent sentiments. For that reason our team opted to remove that category from our model, which did not have a large impact on the amount of data left over for training purposes. We also fine-tuned the pre-processing stage of our code to remove any unnecessary elements from our data, such as punctuation, username tags, stop words, and html links that may have lowered the accuracy of our model. In addition, an "early stop" was implemented, in which the model would stop training after it detected any evidence of overfitting during training.

Finally, to better train the model more data that fell more explicitly into the "Risk of Vaccine" category was annotated and added to the training set in order to give the model a more even distribution of data to train with. In the end the dataset had 475 "Government Conspiracy" tweets, 433 "Medical Freedom" and 418 "Risk of Vaccine" tweets. This step was taken in order to ensure that the model could achieve a higher accuracy in predicting tweets falling in the "Risk of Vaccine" category. The following are the results from that step:

	accuracy	val_accuracy	loss	val_loss
Epoch 1	0.4595	0.7220	1.0443	0.8739

Epoch 2	0.7162	0.7220	0.7760	0.6552
Epoch 3	0.7526	0.7593	0.6140	0.6110
Epoch 4	0.8170	0.7386	0.5196	0.6334
Epoch 5	0.8607	0.7427	0.4042	0.6479

Figure 5.



Figures 6 & 7

	Precision	Recall	F1-score	Support
0	0.80	0.76	0.78	97
1	0.70	0.77	0.73	95
2	0.74	0.65	0.70	49
Accuracy			0.74	241
Macro Avg	0.75	0.73	0.73	241
Weighted Avg	0.75	0.74	0.74	241

Figure 8.

As can be shown in the final results of our model, the accuracy reached an average of 75% over five epochs of training. In the end, we were satisfied with these results, though we did still find that improvements to the model could be made moving forward.

Future Work

The most obvious way to improve this model would be to develop a larger dataset. A dataset with more tweets for all categories in general with more balanced data, especially in currently lacking categories such as Risk of Vaccine. More tweets would also allow for an adequate testing data split without having to worry about obtaining less information in the training phase.

It would also be interesting to explore intent analysis between the 3 categories. That is to compare each of our 3 categories (Conspiracy, Medical Freedom, Risk of Vaccine) against the misinformation, disinformation, and malinformation categories and see which one they align with the most on average.

Additionally, we could use a text explainer to try to obtain explanations into the outcomes behind the classification in order to identify keywords and get a qualitative understanding of the model. This can be done using a tool such as LIME and would help us grasp the workings behind the black-box nature of our model.

Furthermore, we could deploy the model as a web application.This would allow others to test the model and would also be an opportunity to gauge model performance and obtain feedback.

Finally, The twitter data from our model is from April 2021 and it would be interesting to see how it's performance would change when presented with more recent tweets. We could expect a significant difference in performance due to the fast changing nature of the Covid-19 situation around the world.

Conclusion

Our goal was to classify intent behind negative Covid-19 Vaccine tweets using natural language processing. We were able to do the data preparation for this by downloading tweets using the Twitter API and creating a dataset by first, cleaning the data, identifying the major underlying categories to divide it up to, and then annotating the data. From there on, we tokenized the data and implemented the BERT model to perform the classification task and fine tuned the model until an average accuracy of 75% was achieved.

We found that implementing a DistilBert model worked best when working with limited data as we had. Our model was improved significantly by good preprocessing and by training a more balanced dataset. But perhaps, most notably, this model produced satisfactory results indicating that the state-of-the-art BERT model is capable, although not proficient, at nuanced classification tasks such as this, which take place in a specific subdomain - as opposed to a more "straightforward" positive-negative-neutral sentiment analysis task.

Annotated Bibliography

Ai, L., Chen, R., Gong, Z., Guo, J., Hooshmand, S., Yang, Z., & Hirschberg, J. (2021). Exploring New Methods for Identifying False Information and the Intent Behind It on Social Media: COVID-19 Tweets. http://www.cs.columbia.edu/speech/PaperFiles/2021/SocialSens2021_April18.pdf

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). GoEmotions: A dataset of fine-grained emotions. arXiv preprint arXiv:2005.00547.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Muric, G., Wu, Y., & Ferrara, E. (2021). COVID-19 Vaccine Hesitancy on Social Media: Building a Public Twitter Dataset of Anti-vaccine Content, Vaccine Misinformation and Conspiracies. *arXiv preprint arXiv:2105.05134*.https://arxiv.org/abs/2105.05134

Ortiz-Sánchez, E., Velando-Soriano, A., Pradas-Hernández, L., Vargas-Román, K., Gómez-Urquiza, J. L., Cañadas-De la Fuente, G. A., & Albendín-García, L. (2020). Analysis of the Anti-Vaccine Movement in Social Networks: A Systematic Review. *International journal of environmental research and public health*, *17*(15), 5394. https://doi.org/10.3390/ijerph17155394

To, Q. G., To, K. G., Huynh, V.-A. N., Nguyen, N. T. Q., Ngo, D. T. N., Alley, S. J., … Vandelanotte, C. (2021). Applying Machine Learning to Identify Anti-Vaccination Tweets during the COVID-19 Pandemic. *International Journal of Environmental Research and Public Health*, *18*(8), 4069. doi:10.3390/ijerph18084069

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.