Detect Group Bias in COVID-19 Anti-vax Detection Models

Sheng-Tai Huang
shengtai.huang@pitt.edu

ABSTRACT

Due to the prevalence of the anti-vax movement during the pandemic of COVID-19, researchers have been working on developing methods to detect anti-vax articles and prevent them from spreading. However, people sometimes reveal their political stances, gender, or other information that can be related to their identity. If we directly build NLP models on vaccine related articles to detect anti-vax, the models might unwarranted learned word representation related to political stances or gender rather than attitudes toward vaccines. Therefore in this project, I aim at validating whether an anti-vax detection NLP model will learn vaccine related representation or be another political stance or gender detection model.

KEYWORDS

Anti-vax detection, Group disparity, NLP model

ACM Reference Format:
Sheng-Tai Huang. 2022. Detect Group Bias in COVID-19 Anti-vax Detection Models . In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

During the pandemic of COVID-19, although scientists have already developed vaccines to mitigate the spread of virus, due to vaccine hesitancy, there are still many people who decide to postpone getting vaccines or even not to take the vaccines, which prevents the entire society from having herd immunity. Furthermore, the misinformation that exaggerates the negative effects of vaccines and urges people not to take vaccines, known as the anti-vax movement, influences people's willingness to take vaccines. Therefore, researchers have made an effort to understand which factors cause people's anti-vax stance, detect and prevent anti-vax misinformation from spreading. In [6], Roberts et al. investigate several factors such as age, gender, political stances to understand these factors correlation to anti-vax attitude, and the results show that females and Republicans are more likely to become anti-vaxxers. Roberts et al. mention that traditionally females are the main caregivers of the family, and thus might be more likely to participate in the anti-vax movement. Besides, there are also some stereotypes of anti-vaxxers that they are mainly left-wing moms related to the measles outbreak in recent years. Moreover, a survey done by Gallup shows that more and more Republicans have become anti-vaxxers as Figure 1 shown. Hence, the stereotypes and the high prevalence rate of anti-vaxxers in specific groups might make researchers to be unaware of building anti-vax detection model that is unrelated to vaccine attitudes, and [3] already show that gender can be assumed only by word use, so without concerning these potential shortcut, the anti-vax detection model might unwarranted become another political stance or gender detection model and disproportional predict articles from these group as anti-vax. To investigate this issue, this project aims

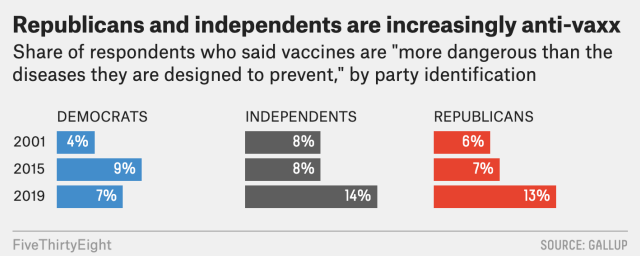

Figure 1: Proportion of anti-vaxxers in each political stance (Source: "Republicans Aren't New To The Anti-Vaxx Movement", by Kaleigh Rogers, FiveThirtyEight, URL: https://fivethirtyeight.com/features/republicans-arent-new-to-the-anti-vaxx-movement/)

at answering the following two research questions:

- Are anti-vax detection NLP models just another stance or gender detection models?
- If some groups tend to be mistakenly labeled as anti-vax, what causes the model to make the wrong predictions?

2 METHOD

In this project, I aim at discovering the group bias within an anti-vax detection NLP model. However, since there is no pre-trained NLP model specific for COVID-19 vaccine, I use a pre-trained NLP model that is based on Twitter data about COVID-19, called COVID-Twitter-BERT [4], and a Twitter data that has labels whether each post is anti-vax or not, called ANTiVax [2], to fine-tune the COVID-Twitter-BERT model. The model evaluation is not only based on the labeled data, but also collected Facebook posts by using Crowdtangle API. The reason fro using Facebook posts since there are many public groups on Facebook, which can assume the identity of members without the help of using other machine learning algorithms. Subsequently, I manually label the collected data with anti-vax labels, and since all data have an imbalance class distribution of anti-vax stances, accuracy, precision, recall, and F1-score are all considered to compare the group disparity. After model evaluation, since I want to understand what text representation the model has learned, the BERTopic [1], which uses the embeddings from transformer models to build topic models, to see whether there are topics or the important words in topics are unrelated to vaccine discussion but political stances or gender. However, originally the

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Table 1: Keywords for searching groups

Group	Keywords
Conservative/Republican	conservative, republican
Libertarian/Democrat	libertarian, liberal, democrat
Mom	mother, mom, mommy, mama
Dad	father, dad, daddy, papa

BERTopic uses sentence embeddings from sentence transformers to build topic models, which cannot directly apply to COVID-Twitter BERT since it only learns word embeddings. Therefore, the word embeddings of each sentence are aggregated and averaged as the sentence embeddings, and follow the data pre-processing steps in [1], the stop words are removed from sentences.

3 DATA

In this project, two data sets are used, one is for fine-tuning and the other serves as test data to evaluate the COVID-19 anti-vax detection NLP model. The data for fine-tuning is from [2], originally it has a total 15,073 samples, with 9,322 not anti-vax and 5,751 anti-vax samples. However, since some accounts were banned by Twitter or users removed their posts, some posts are inaccessible by Twitter API, , and thus there are only 12,692 samples in total, with 8,297 not anti-vax and 4,395 anti-vax samples. Since Twitter users can only assume their identity based on their pictures (which might not be themselves) and account names (which might be pseudonyms) as [5] has done. Therefore, I choose Facebook to collect my data since users' identity such as political stances and gender can be easily assumed based on the groups they belong to predict their group identity. To search for vaccine related posts, I use Crowdtangle API to search with two keywords, "vaccin" and "vax", from Jan. 2020 when the first reported cases of COVID-19 occurred to Feb. 2022. In my experiments, due to the existing stereotypes or results of previous papers, I compare group disparity of Conservatives/Republicans vs. Libertarians/Democrats and moms vs. dads. These groups are searched by the following keywords as Table 1 shown. However, the search of public groups might contain posts from different countries, which might have different vaccine policies thus be incomparable for group disparity. For example, Canada and the UK also have Conservative Party, which increases the number of posts from Conservatives/Republicans, and thus posts will be removed if they contain the words "Trudeau", "Canada", "Canadian", "UK".

The data pre-processing follows both [4] and [2]. First, external links and tagged account names are removed since the vocabulary of the tokenizer cannot handle the diverse external links, and tagged account names might potentially reveal authors' political stances or gender, which might unwarranted become model bias. Second, emojis are replaced with ASCII representations, which can be considered tokens while training and provide more information for predicting vaccine attitude. Finally, posts only with photos are not considered since the NLP model is used in this project, which cannot handle images, and posts only with links are not considered since it is impossible to understand authors' intention of sharing links without any comments. Before annotating posts whether they are anti-vax or not anti-vax, it is important to understand the official vaccine stance of the two main parties. Both the Republican Party and Democratic Party encourage their supporters to take vaccines. However, the Republican Party officially opposes the vaccine mandate, as in Republican National Committee (RNC) Statement on vaccine mandate – "While I am pro-vaccine, the Biden administration does not have the authority to force hardworking Americans to choose between being vaccinated and providing for their families. That's why the RNC is suing the Biden administration over this unlawful vaccine mandate and will maintain every legal option to fight this authoritarian overreach." Therefore, people especially for Republicans might have the following statements: "I'm vaccinated, but I against vaccine mandate", "I'm pro-vax, but I against vaccine mandate", so label anti vaccine mandate as anti-vax might bias toward Conservatives/Republicans although some anti vaccine mandate people are also anti-vaxxers. Therefore, anti vaccine mandate posts are not considered in my experiment. After I manually read through all posts and label their vaccine attitude, Table 2 shows the class distribution of the four groups. We can see that Conservatives/Republicans there are more likely to post anti-vax articles. However, since there are too few posts from dads' groups, rather than compare between moms and dads, the group disparity is compared between mom-specific groups and other two not mom-specific political groups in the experiment.

Table 2: Class distribution of each group

Group	Not Anti-vax	Anti-vax
Conservative/Republican	41 (28.5%)	103 (71.5%)
Libertarian/Democrat	46 (90.2%)	5 (9.8%)
Mom	27 (77.1%)	8 (22.9%)
Dad	1 (50%)	1 (50%)

Table 3: Fine-tune

	Training	Validation
Accuracy	0.9919	0.9882
Precision	0.9934	0.9875
Recall	0.9832	0.9786
F1-score	0.9883	0.9831

4 EXPERIMENT AND DISCUSSION

In my experiment, the anti-vax NLP model is a binary classifier: not anti-vax and anti-vax. However each group has imbalanced samples on anti-vax and not anti-vax, and thus not only accuracy but also precision, recall, F1-score are considered. For initial experiments, I use all samples in ANTiVax data [2] to fine-tune the entire COVID-Twitter-BERT for learning vaccine related representations. However, although the fine-tuned model can reach high model performance as Table 3 shown, it performs worse on the collected data. Therefore, I also split my collected Facebook data as training and test set to fine-tune and evaluate the BERT model, but due to the limited number of training samples, even different random seeds can change the model prediction, and the performance is still not

Table 4: Metrics between Conservative/Republican (abbr. as Con/Rep), Libertarian/Democrat (abbr. as Lib/Dem), and Moms

Group	Con/Rep	Lib/Dem	Mom
Accuracy	0.5278	0.7451	0.7222
Precision	0.8889	0.1	0
Recall	0.3884	0.2	0
F1-score	0.5405	0.1333	0

acceptable thus not reported in this section. As previously mentioned, the terrible performance might be because the model learns vaccine unrelated representations to predict anti-vax labels. Thus, the BERTopic is applied to the fine-tuned COVID-Twitter-BERT and the training data of ANTiVax is used to draw scatter plots of posts and derive the most important words of each topic. As Figure 2 shows, although there are five topics detected, most posts belong to topic 4 and 5, which might be a good sign that posts are clustered by their labels. Furthermore, the following list of top 10 words for each topic shows that the top most important words are meaningful enough for people to recognize the subjects of topics, and Topic 4 seems to represent anti-vax and Topic 5 for not anti-vax.

- Topic 1 No Vaccine Mandate: free, nomandates, noforced-covidvaccines, belong, unconstitutional, unethical, tracked, mandates, bodies, choose
- Topic 2 Influence of Vaccine Mandate: incarceration, licenses, ubiquitous, isolation, pointless, universal, movement, worst, passports, completely
- Topic 3 Vaccine is gene therapy: excuse, herd, using, immunity, kids, therapy, gene, voodoo, growth, tea
- Topic 4 Experimental Vaccine: experimental, gene, therapy, depopulation, people, covid, vaccines, mrna, virus, untested
- Topic 5 Get the vaccine: got, dose, vaccinated, worry, today, covid, second, just, don, grateful

However as Figure 3 shown, both non anti-vax and anti-vax posts are mixed up in both topic 4 and 5, so it requires further investigation to understand whether the BERT model does not learn vaccine related representations or just I aggregate word embeddings as sentence embeddings in a wrong way.

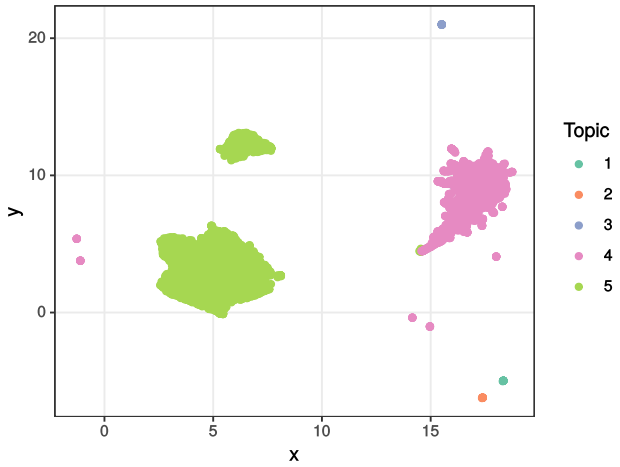

Figure 2: Scatter plot of training samples

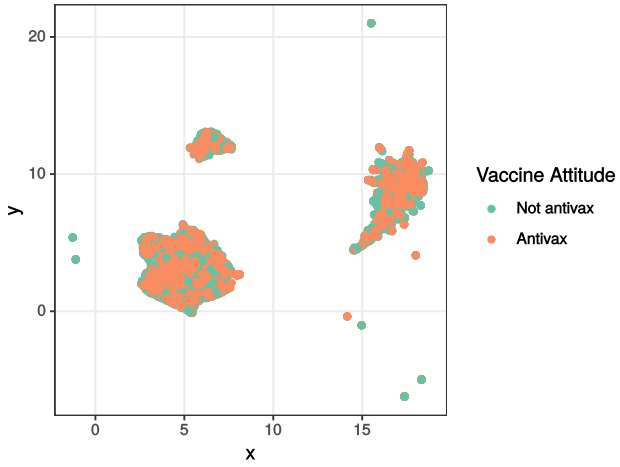

Figure 3: Density Plot of Tags

5 CONCLUSION

From the distribution of collected data shows that Conservatives/Republicans are more likely to post anti-vax articles compared to Libertarians/Democrats, and although it is not comparable between moms' and dads' groups, there are more moms discuss about vaccines, so by seeing the number of posts it seems to have more moms compared to dads join the anti-vax movement. In the experiment of group disparity, Conservatives/Republicans vs. Libertarian/Democrats and moms vs. not mom specific groups are compared, but the fine-tuned model does not fit well with the collected Facebook data, and thus I cannot make an exact conclusion about what causes the group disparity. The potential problems might be:

- The fine-tuned model learned unrelated text representations.
- My own bias on labeling anti-vax data.
- Existing group bias in the labeled ANTiVax data
- Different maximum sequence length of Twitter and Facebook which makes the BERT models are able to handle Twitter data rather than Facebook

Consequently, for future research, rather than using different data sources, I should use the labeled ANTiVax data and follows [5] to detect users' political stances and gender to reproduce the entire experiment pipeline to understand whether the group disparity really exists.

REFERENCES

[1] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794* (2022).
[2] Kadhim Hayawi, Sakib Shahriar, Mohamed Adel Serhani, Ikbal Taleb, and Sujith Samuel Mathew. 2022. ANTi-Vax: a novel Twitter dataset for COVID-19 vaccine misinformation detection. *Public health* 203 (2022), 23–30.
[3] Zachary Miller, Brian Dickinson, and Wei Hu. 2012. Gender prediction on twitter using stream algorithms with n-gram character features. (2012).

[4] Martin Müller, Marcel Salathé, and Per E Kummervold. 2020. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *arXiv preprint arXiv:2005.07503* (2020).
[5] Ignacio Ojea Quintana, Marc Cheong, Mark Alfano, Ritsaart Reimann, and Colin Klein. 2022. Automated clustering of COVID-19 anti-vaccine discourse on Twitter. *arXiv preprint arXiv:2203.01549* (2022).
[6] Hannah A Roberts, D Angus Clark, Claire Kalina, Carter Sherman, Sarah Brislin, Mary M Heitzeg, and Brian M Hicks. 2022. To vax or not to vax: Predictors of anti-vax attitudes and COVID-19 vaccine hesitancy prior to widespread vaccine availability. *Plos one* 17, 2 (2022), e0264019.