


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



Presenter: Sheng-Tai Huang

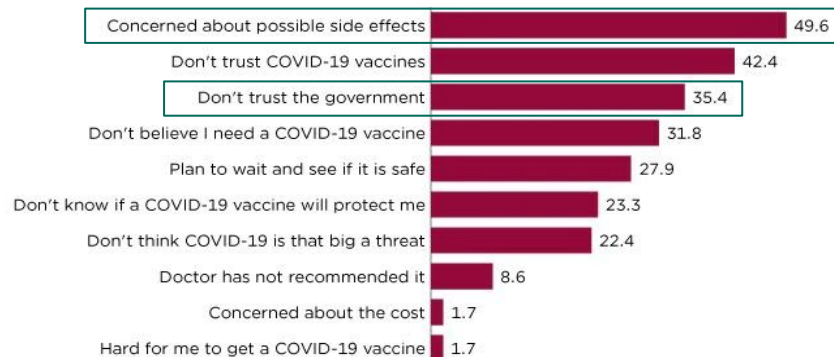


Motivation

COVID-19 anti-vax stances might be related to specific groups of people, for example:

- “Concerned about possible side effects” can be related to family caregivers
- “Don’t trust the government” can be related to supporters of the opposition party

Why Adults 18 and Over Did Not Get COVID-19 Vaccine
(In percent)



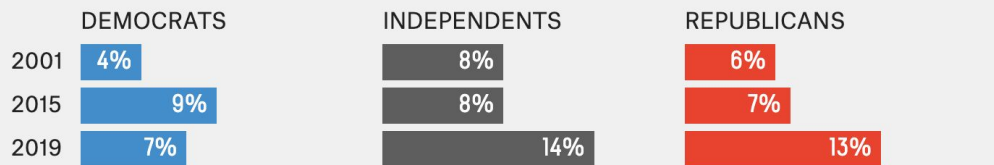
Note: Responses do not sum to 100 as respondents could select more than one reason.
Source: U.S. Census Bureau, Household Pulse Survey Week 40 (December 1-13, 2021).

Motivation (cont.)

During the pandemic of COVID-19, the **Conservatives/Republicans** are more likely to be anti-vaxxers compared to Liberterians/Democrats.

Republicans and independents are increasingly anti-vaxx

Share of respondents who said vaccines are "more dangerous than the diseases they are designed to prevent," by party identification



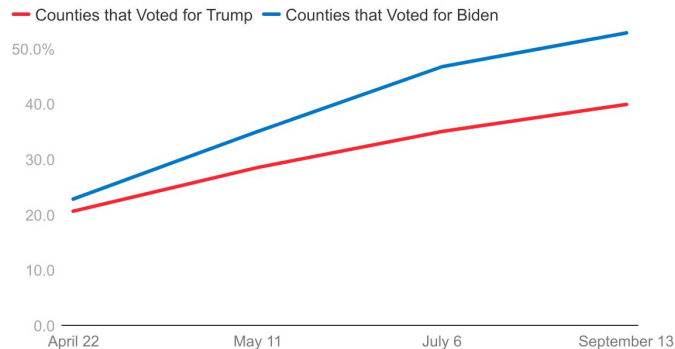
FiveThirtyEight

SOURCE: GALLUP

Source: <https://fivethirtyeight.com/features/republicans-arent-new-to-the-anti-vaxx-movement/>

Figure 1

Vaccination Rates in Counties that Voted for Biden and Counties that Voted for Trump, April - September 2021



NOTE: Data are share fully vaccinated.
SOURCE: KFF analysis of CDC's COVID-19 Integrated County data.

Motivation (cont.)

Another common stereotype of the the anti-vax movement in America is that people who are **left-wing, coastal, white, wealthy moms** might be more likely to become anti-vaxxers.

Therefore, potential group bias might exist in COVID-19 anti-vax detection classifiers for:

- Moms vs. Dads
- Conservatives/Republicans vs. Liberterians/Democrats

Research Questions

Many researches use NLP models to detect anti-vax posts, misinformation, etc.

However, there are also stance or gender detection NLP models.

RQ1: Are anti-vax detection NLP models just another stance or gender detection models?

RQ2: If some groups tend to be mistakenly labeled as anti-vax, what causes the model to make the wrong predictions?

Model

- Anti-vax classifier: Fine-tuning COVID-Twitter-BERT, which is based on BERT-LARGE
- BERTopic: Extract the learned embeddings from the transformer, and do topic modeling to observe

Data

- Data for fine-tuning: Labeled ANTiVax data (with total 12,692 samples, 8,297 not anti-vax and 4,395 anti-vax samples)
- My collected data: Facebook data search with “vaccin” and “vax” by using Crowdtangle, Jan. 2020 ~ Feb. 2022, only consider public groups and verified profiles

Data Preprocessing

- External links, account names (@user) are removed.
- Emojis are replaced with ASCII representations.
- Posts only with links or photos are not considered.

How to Label Group Identity

Identities of authors are assumed by the groups they belong to

- Authors belong to “Real Jewish Conservatives” are regarded as Conservatives/Republicans
- Groups with name such as “NVR: NEVER vote republican” are removed

Challenge of Data Labeling

It is difficult to classify anti vaccine mandate posts as anti-vax or not.

For example, some people said that

- “I’m vaccinated, but I against vaccine mandate”
- “I’m pro vax, but I against vaccine mandate”



Anti-vaccination activists participate in a rally after a Defeat The Mandates march at the Lincoln Memorial. Alex Wong—Getty Images (Source: TIME)

Challenge of Data Labeling

Also officially both **Democratic Party** and **Republican Party** support COVID-19 vaccines, but **Republican Party** is against the 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**.”

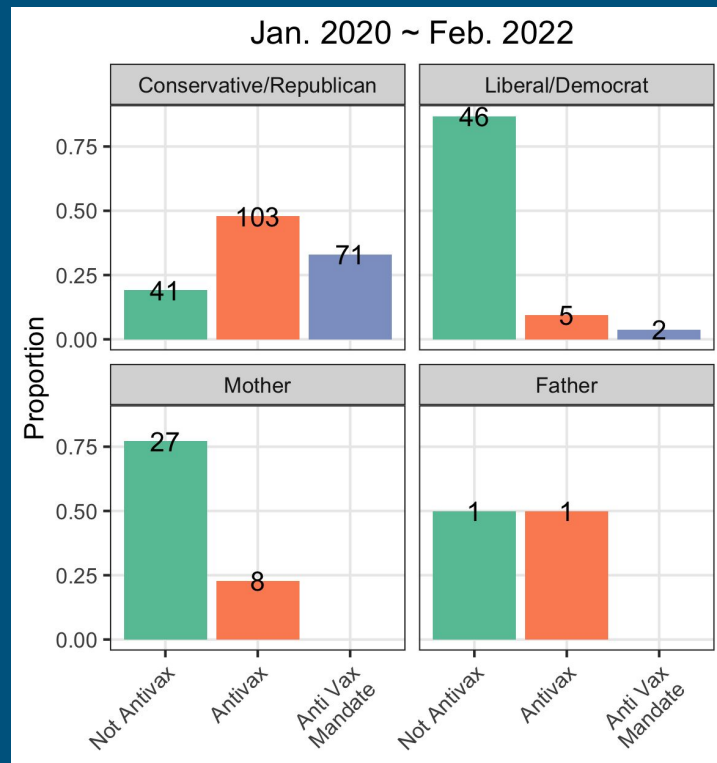
– RNC Statement on Vaccine Mandate

Therefore, anti vaccine mandate is considered a different class and removed from the experiment or label them as anti-vax might inject my own biases.

Facebook Data Distribution

There are more likely Conservatives/Republicans to post anti-vax articles.

However, since there are too few posts from dad's groups, group disparity between moms and dads is not considered in the rest of the experiment.



Metrics

The anti-vax NLP model is a binary classifier: not anti-vax and anti-vax.

However, both political stance groups are imbalanced, the following metrics are considered:

- Accuracy
- Precision
- Recall
- F1-score

Fine-tune Results

The COVID-Twitter-BERT can almost accurately predict the anti-vax labels.

	Training	Validation
Accuracy	0.99192	0.98818
Precision	0.99337	0.98751
Recall	0.98317	0.97863
F1-score	0.98825	0.98305

Results on Facebook Data

The classifier tends to mistakenly predict on the Liberterian/Democrat group.

	Conservative/Republican	Liberterian/Democrat
Accuracy	0.52778	0.74510
Precision	0.88889	0.1
Recall	0.38835	0.2
F1-score	0.54054	0.13333

Summary of Results

Based on the experience of COMPAS, Conservative/Republican group is more likely to be predict with more anti-vax posts.

However, in the experiment, Liberterian/Democrat is more likely to be predict with more anti-vax posts. Therefore, I use BERTopic to extract the learned representations of the NLP model to see what is happened.

How to Use BERTopic

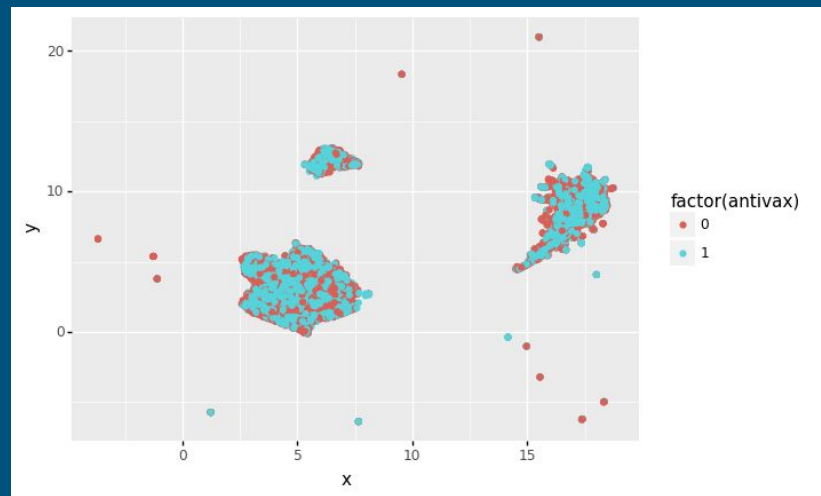
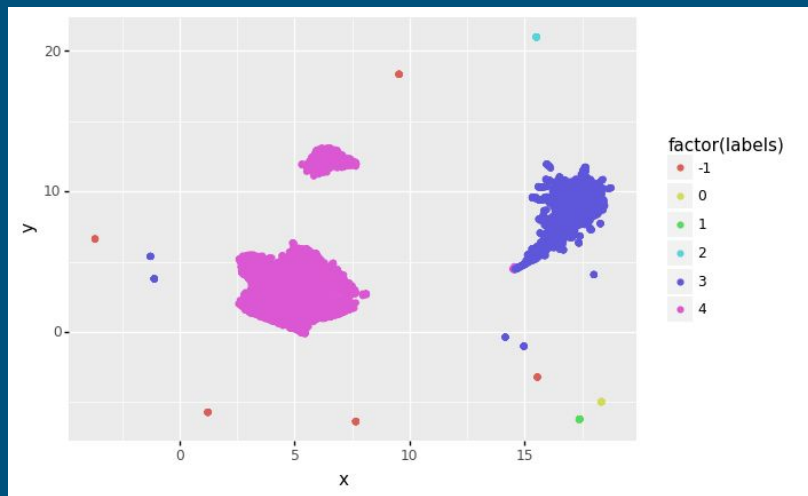
BERTopic uses sentence embeddings from sentence transformers to do topic models, which cannot directly apply to COVID-Twitter BERT (only has word embeddings)

How to apply BERTopic to general BERT?

- Remove stopwords
- Average word embeddings in the same posts.

Results of BERTopic

Five topics are detected for the COVID-Twitter-BERT.



Top 10 Words of Each Topic

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

Take-away

There are more moms' groups and moms are more likely to discuss about vaccines, so it seems to have more moms compared to dads join the anti-vax movement.

From the collected data, Conservatives/Republicans are more likely to post anti-vax articles compared to Liberterian/Democrat.

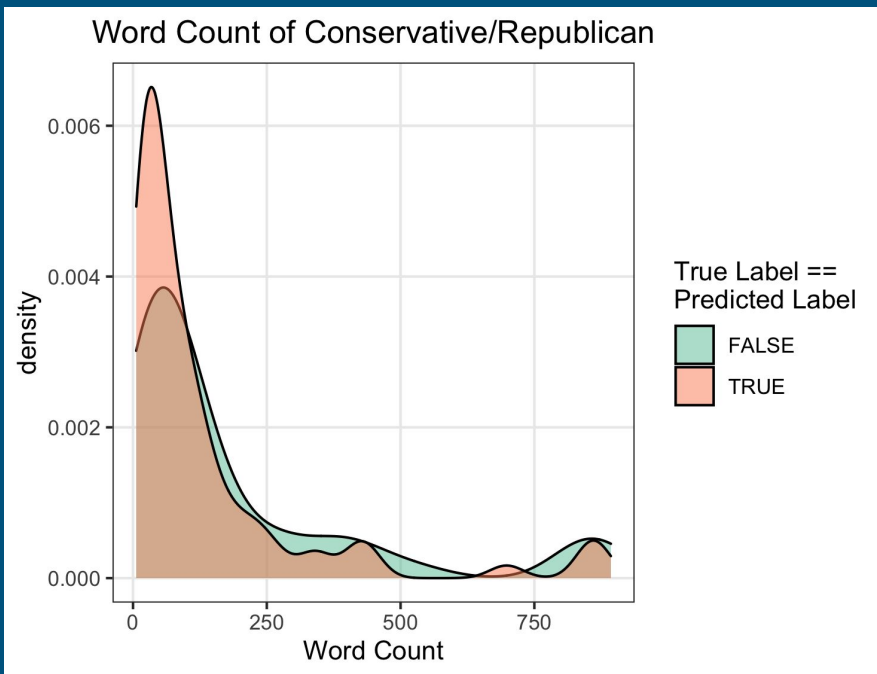
Potential problems:

- Learned unrelated representations during fine-tuning (Shortcut learning)
- Group bias in the fine-tuning data

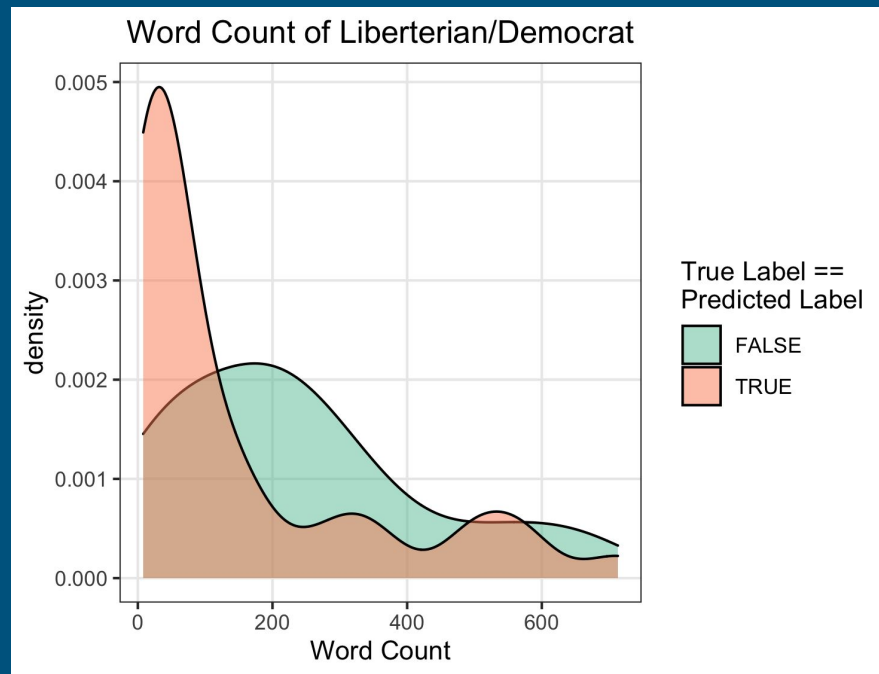
Thanks for listening.

Density of Word Count

Wilcoxon test: p-value = 0.08171



Wilcoxon test: p-value = 0.05437



Project Goal

- Investigating whether a COVID-19 anti-vax detection classifier will have different model performance on different groups of people (e.g., political stances, gender)
- If the group disparity exists, what cause the group disparity