COVID19 Vaccine Twitter Posts: Data Labeling and Text Classification

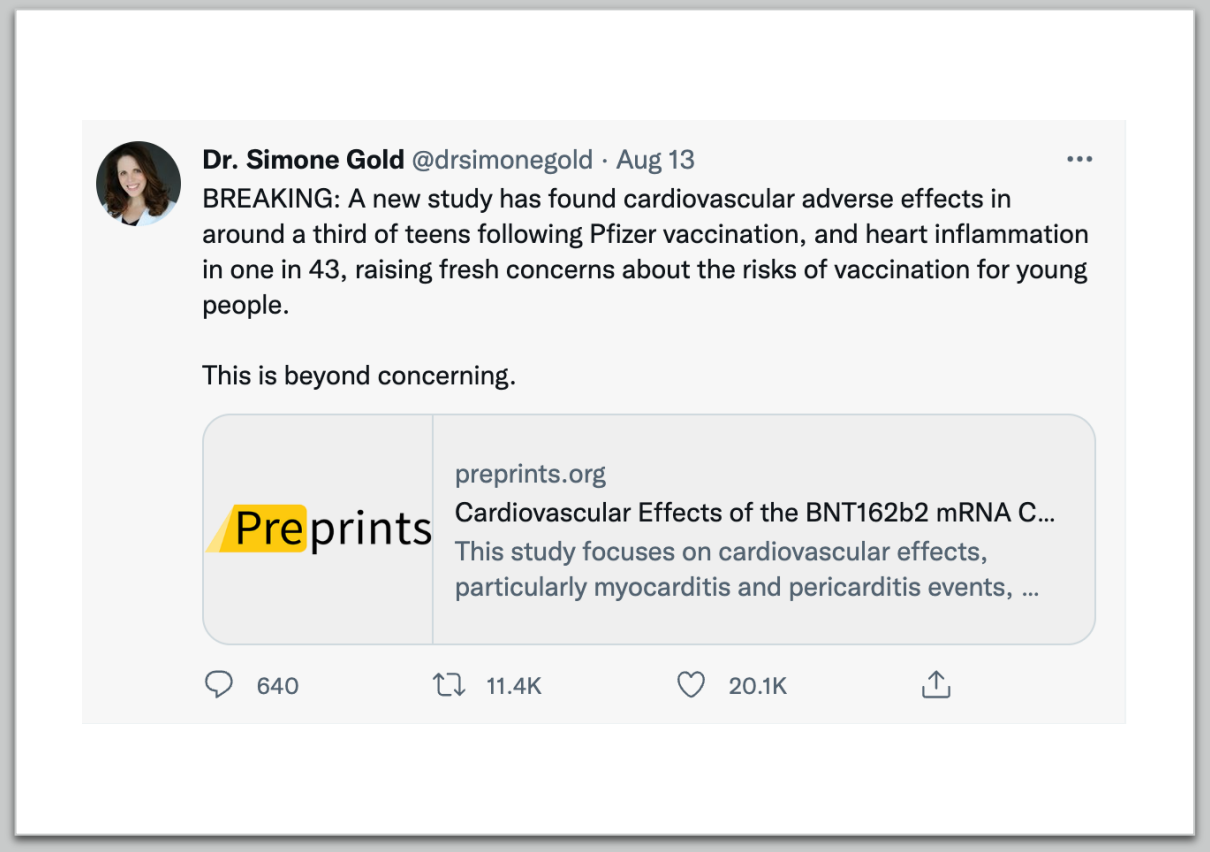
Rhoeun Park

dsir-fx-222

Problem Statement

SNS data serve as great resources for public opinions on different topics, don't have necessary y-labels to predict on.

The goal of this project is to find out an optimal labeling method for twitter posts regarding COVID19 vaccines to classify them into "pro-vaccination" vs "anti-vaccination".

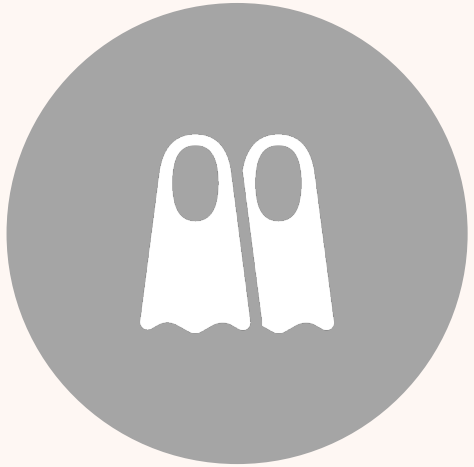


Anti-vax vs Pro-vax

Labeling Methods

HASHTAG –

CROWDSOURCING

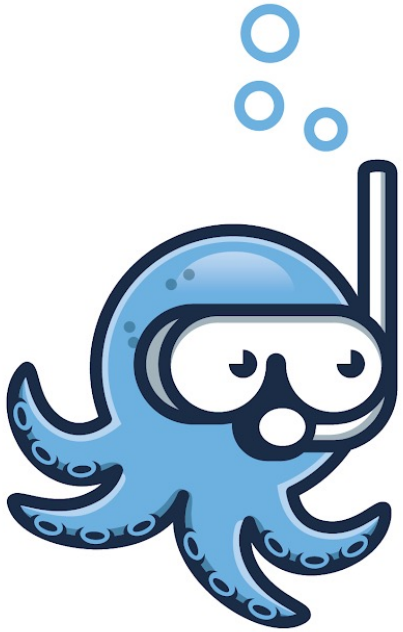


SNORKEL-

DATA LABELING PLATFORM

Snorkel.ai

- Labeling tool for unlabeled data
- Sets of rough rules that help classify between the texts
- Take a vote on the majority!
- Takes weight into consideration



snorkel

Datasets

- SNScrape API
- 2 main datasets

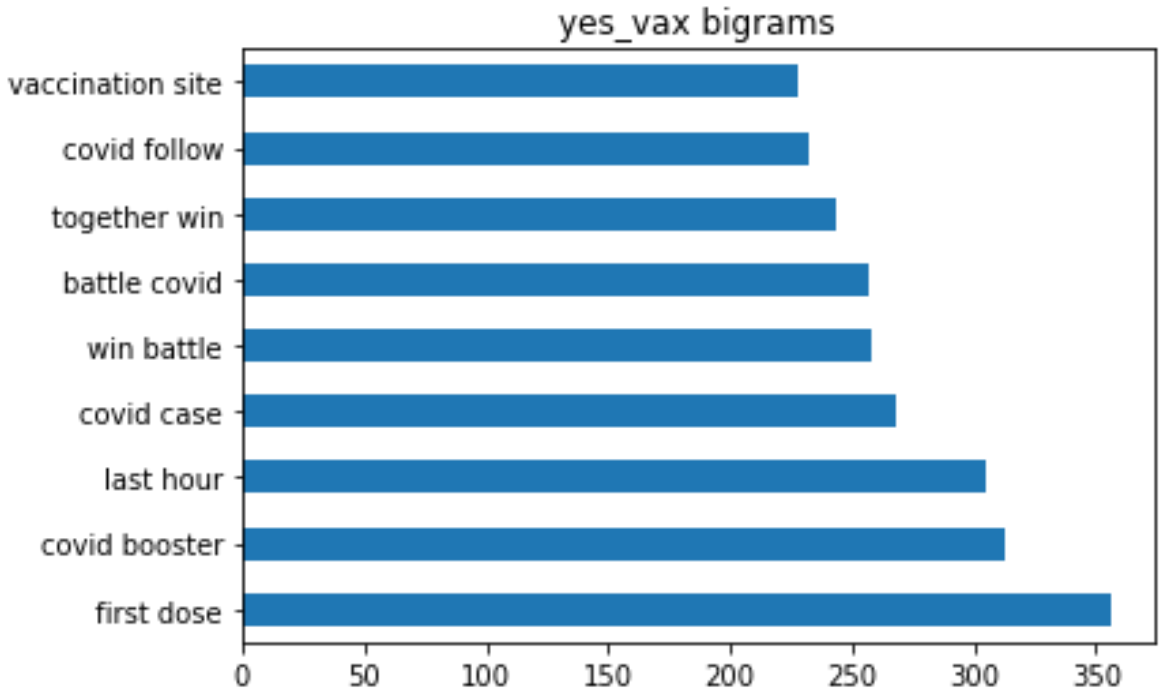
- Unlabeled (data scraped just with general covid19 vaccine topic)

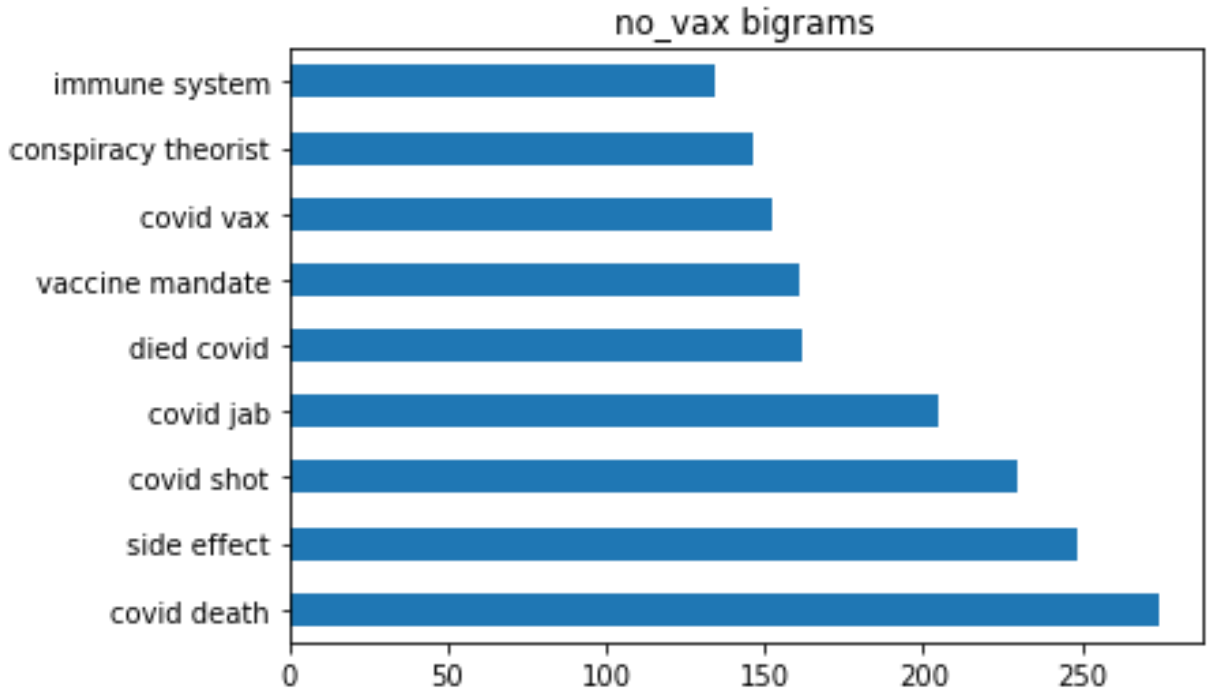
- Hashtag labeled (data scraped based on parameters

Pro_Vax	n
Unite2FightCorona	1000
StaySafe	679
GetVaccinated	473
Vaccinated	417
LargestVaccinationDrive	381
healthcare	360
COVIDisAirborne	337
COVIDAppropriateBehaviour	332

Anti_Vax	n
NoVaccineMandates	1060
InformedConsent	748
MyBodyChoice	545
NoVaccinePassports	528
VaccineSideEffects	513
MedicalFreedom	503
IDoNotConsent	484
VaccineInjury	419

Hashtag Labeled Data – Other Hashtags





Hashtag Labeled - Bigrams

Snorkel labeling function

- Keywords
- Sentiment
- subjectivity
- Hashtags
- Profanity words

Snorkel Results

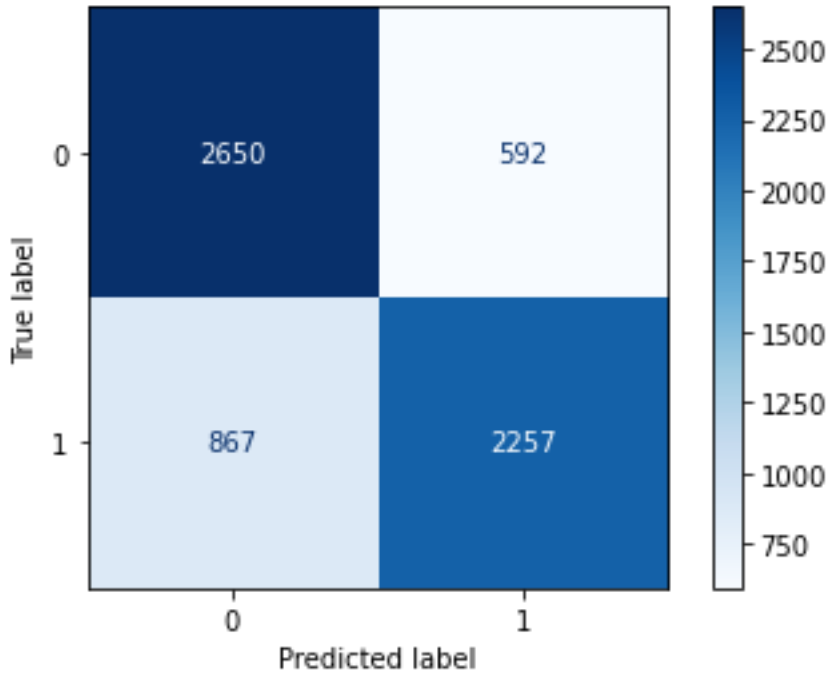
development set accuracy = 59.2%

hashtag subset accuracy = 61.5%

Coverage:
75,644 out of 111,959 rows

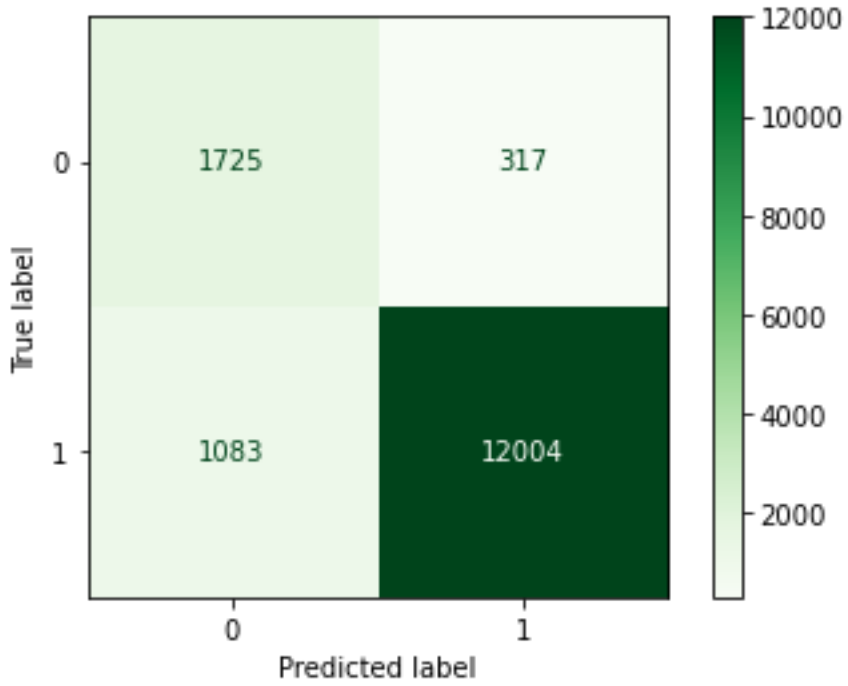
label distribution:
pro-vax (0.86), anti-vax (0.14)

Model results – Hashtag Labeled



test	precision	recall	f1-score
0	0.75	0.82	0.78
1	0.79	0.72	0.76
accuracy			0.77

Model results – Snorkel Labeled



test	precision	recall	f1-score
0	0.61	0.84	0.78
1	0.97	0.92	0.94
accuracy			0.91

Conclusion & next steps

- Snorkel Labeled data is better in terms of accuracy, recall, and precision
- Snorkel model integrity could be questionable
- Hashtag labeling is easier, but limited
- **Next step:**
  - compare across the two models
  - run hashtag labeled against unlabeled dataset