Predicting Protest State Responses

By Denise Macias, Jose Delgadillo & Rhoeun Park

Problem Statement

Can we distinguish between protests that will lead to a negative or non-negative state response?

State violence against protesting civilians: A global comparison

The killings of George Floyd and Breonna Taylor by police in 2020 ignited anti-police brutality and anti-racism protests across the U.S. and the world. For many, the murders represented a shift in awareness of police brutality as an issue—between 2015 and 2020, there was a marked increase in U.S. adults who believed that police violence was a serious problem.

As protesters took to the streets, another shift was occurring: increased police violence against protesting civilians. Over the past several years, the use of excessive force by police and military against protesters—both domestically and globally—has steadily increased, according to a statement from the United Nations Human Rights Office. This trend includes violence against journalists covering protests and has made conditions for protesting more dangerous. The U.S. nearly tops the list for most incidents of state violence against protesters, placing sixth among all countries.

Source: https://stacker.com/stories/40894/state-violence-against-protesting-civilians-global-comparison

Dataset

Mass Mobilization Protest Data *from Harvard Dataverse*

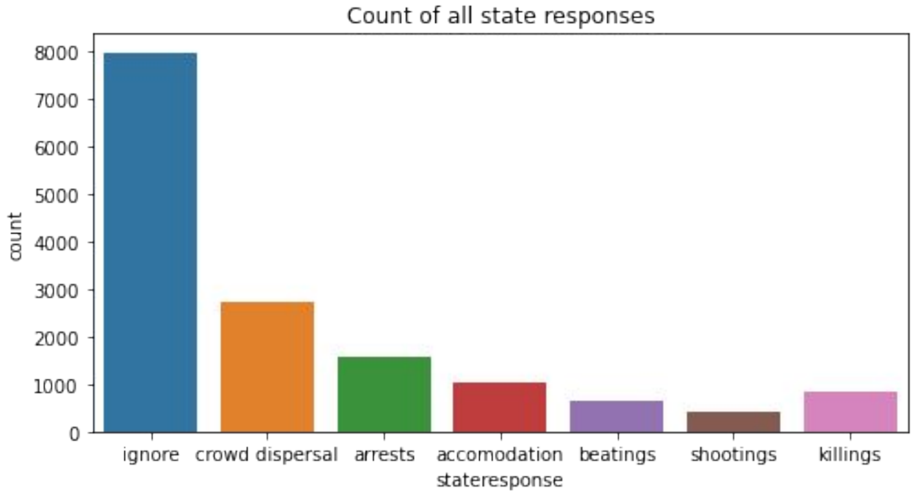
- Protests from 162 countries between 1990 and March 2020
- 15198 instances of mass mobilization events
- 31 features, 1 notes column with open-source text data

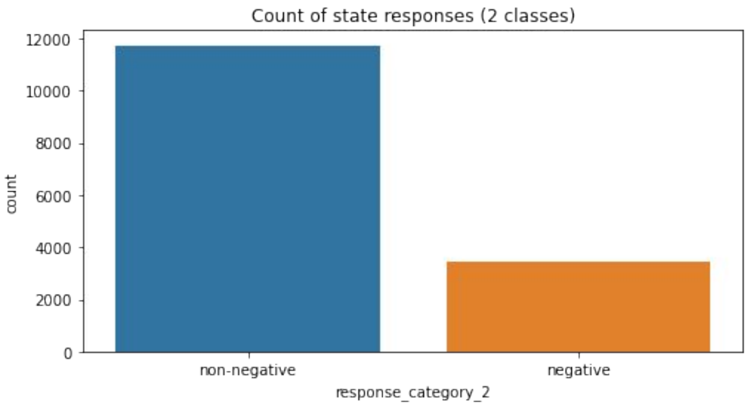
* Notes were taken after the event happened

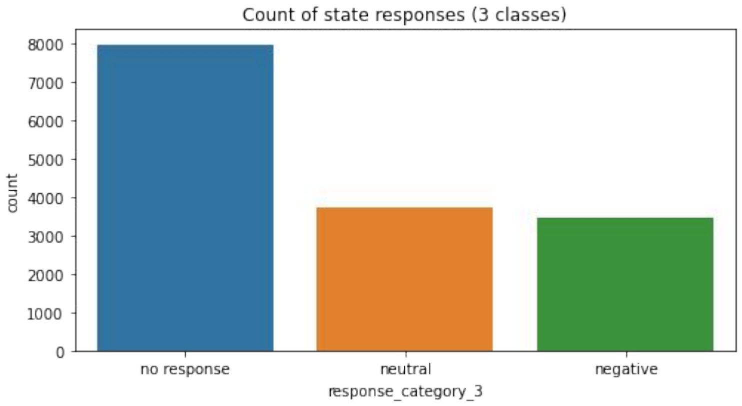
Data Cleaning - Categorical Features

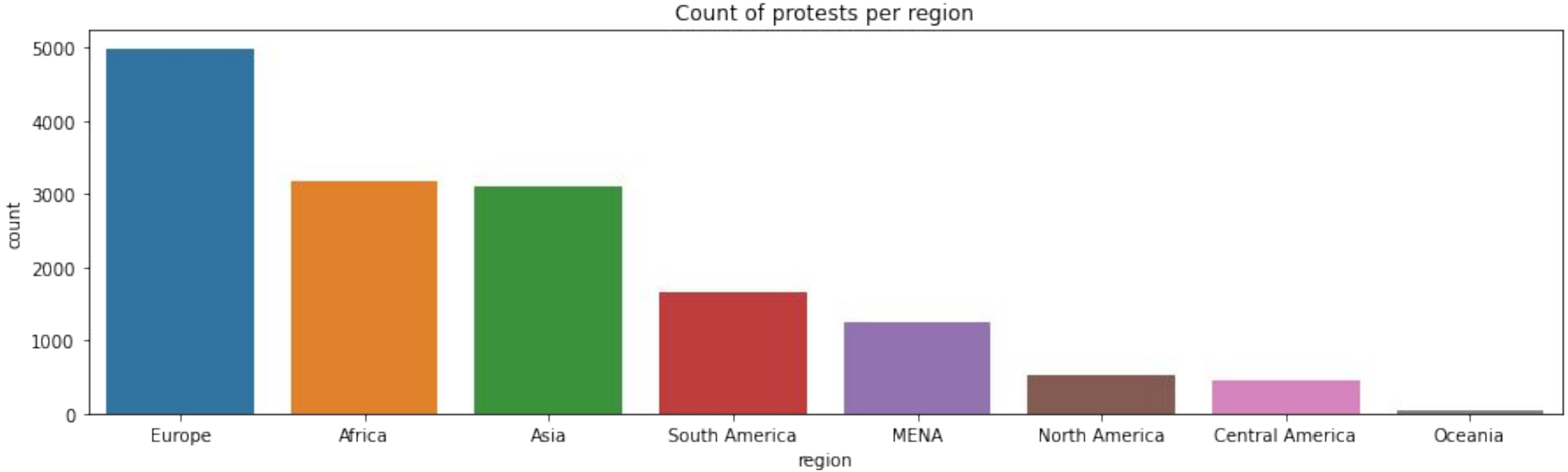
- The original dataset consisted of mainly categorical features
- Cleaning of the target variable
  - 7 state response columns = Up to 7 state responses recorded per protest
  - Target variable class groupings:
    - 2 classes → negative & non-negative
    - 3 classes → negative, neutral & no response

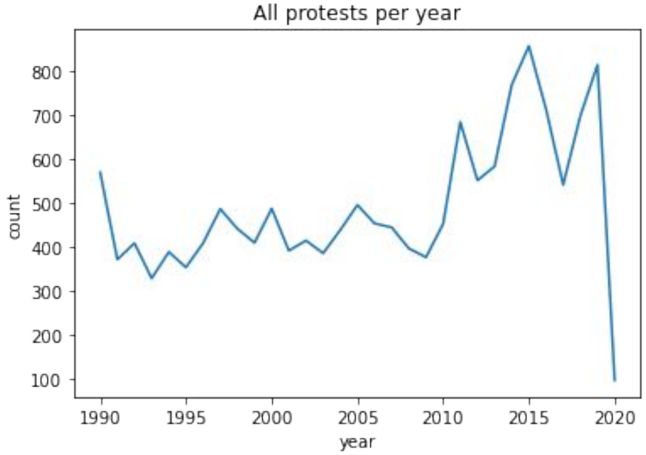
EDA - Categorical Features



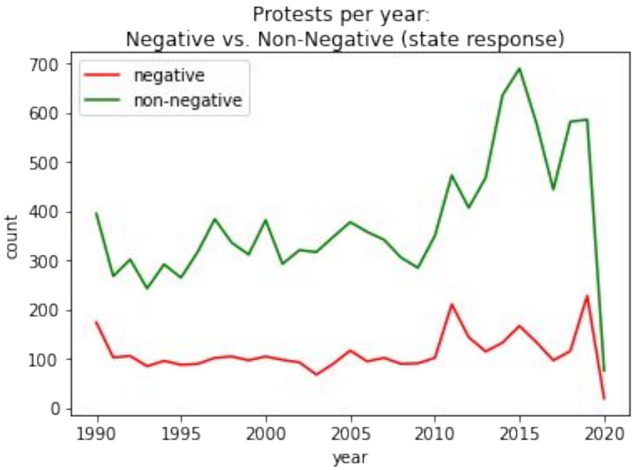




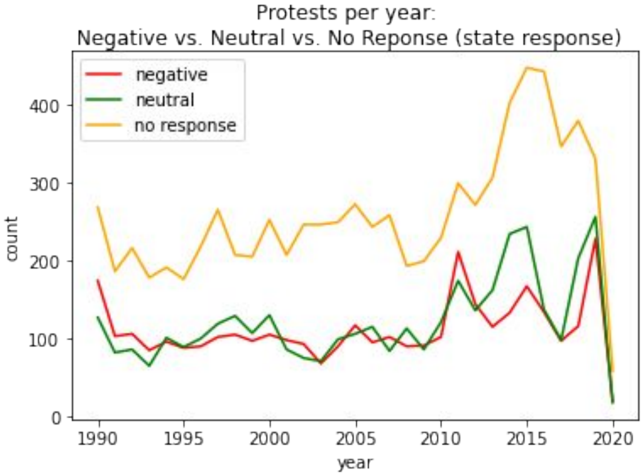

Count of protests per region
count
0
1000
2000
3000
4000
5000
Europe
Africa
Asia
South America
MENA
North America
Central America
Oceania
region



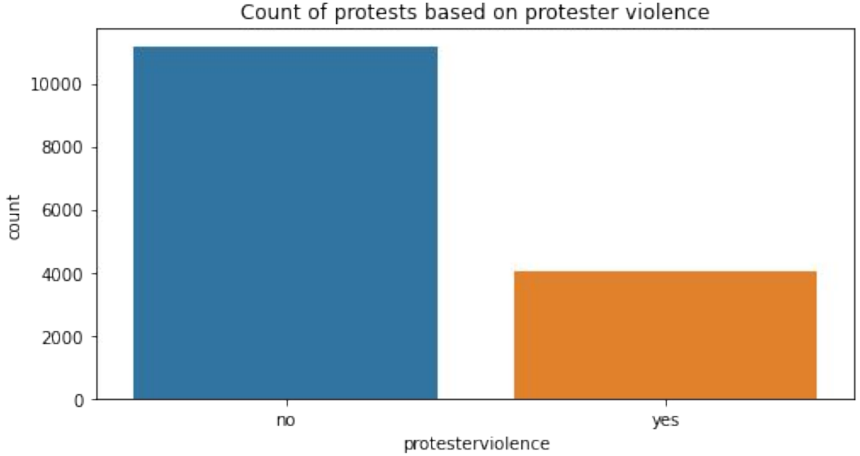
All protests per year
count
year



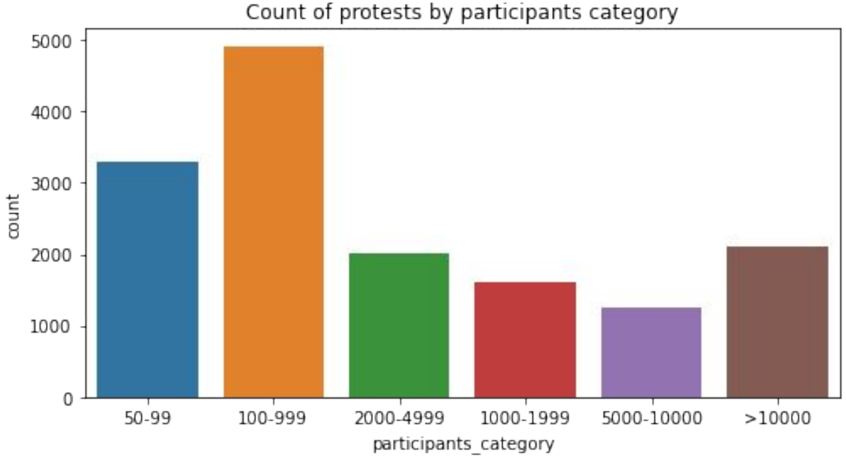
Protests per year:
Negative vs. Non-Negative (state response)
negative
non-negative
count
year



Protests per year:
Negative vs. Neutral vs. No Reponse (state response)
negative
neutral
no response
count
year

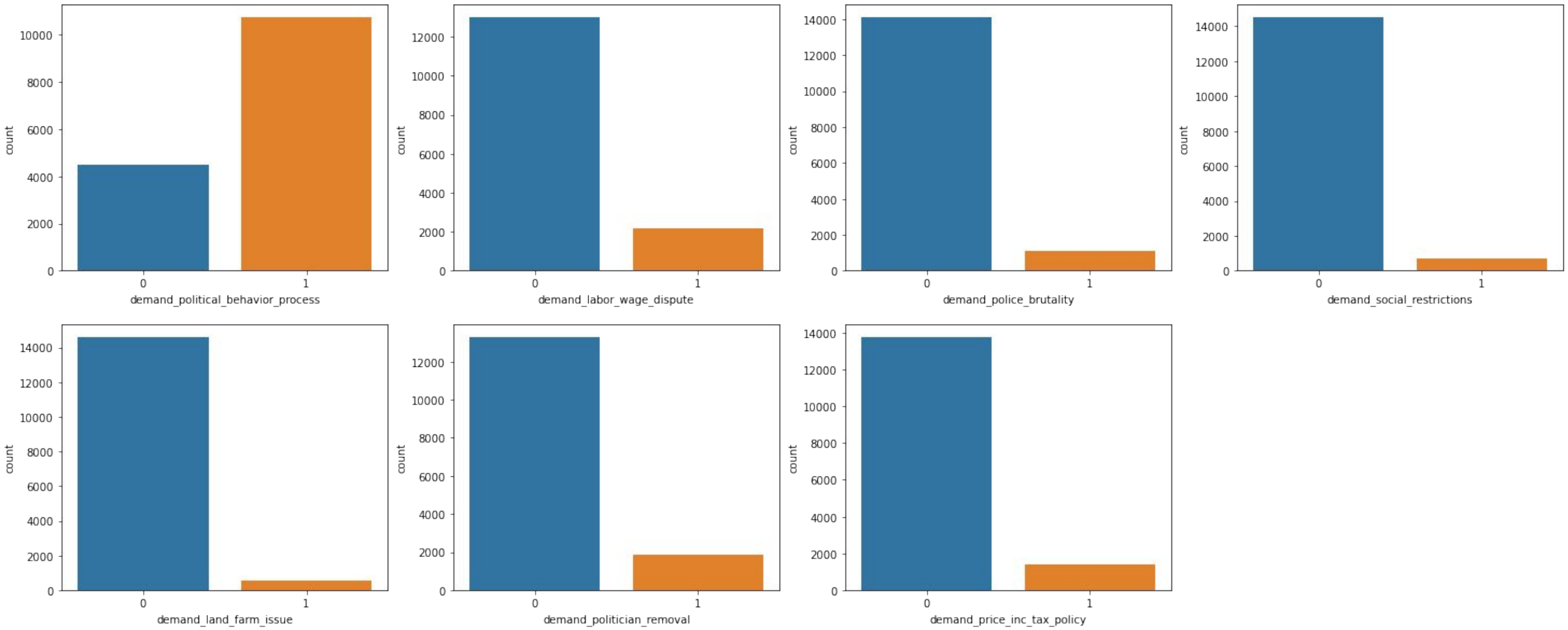


Count of protests based on protester violence
count
10000
8000
6000
4000
2000
0
no
yes
protesterviolence

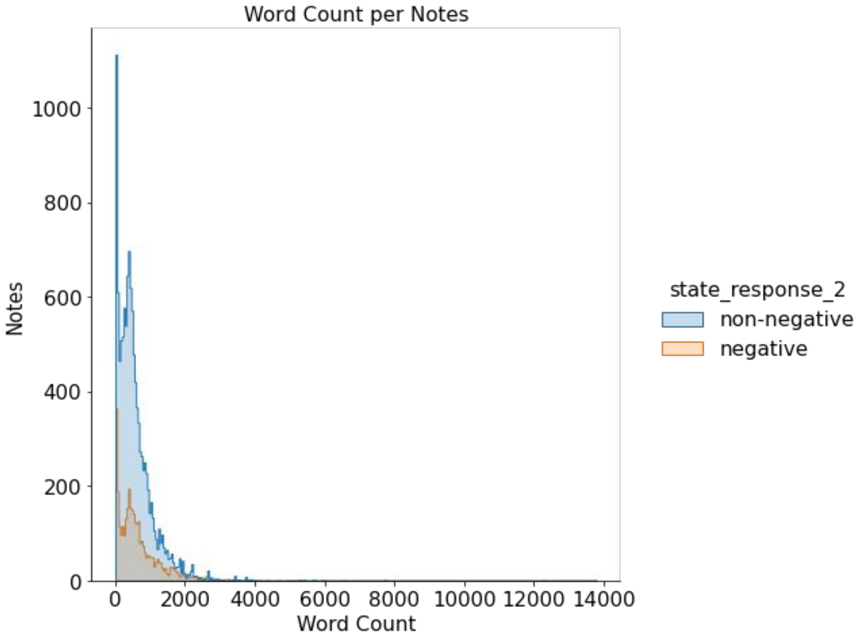


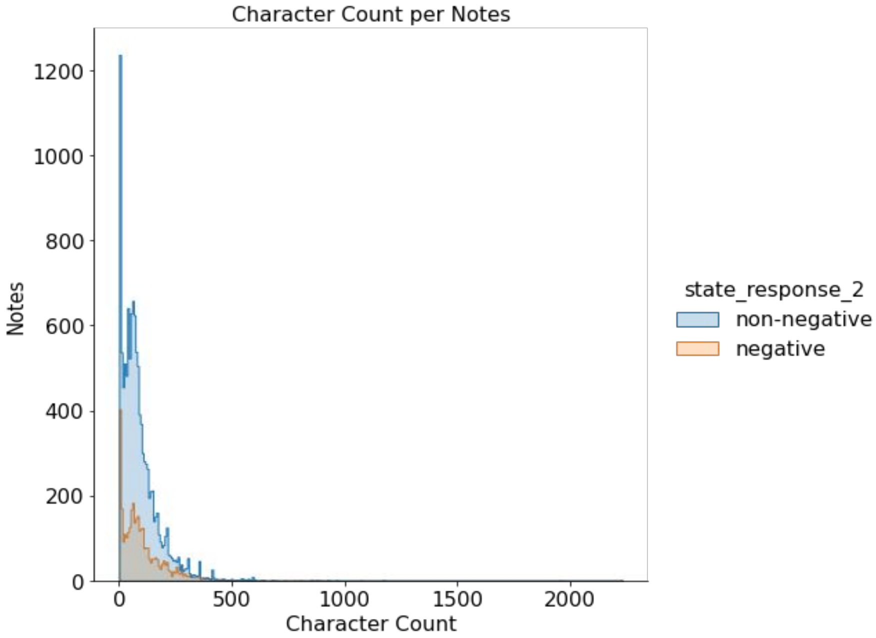
Count of protests by participants category
count
5000
4000
3000
2000
1000
0
50-99
100-999
2000-4999
1000-1999
5000-10000
>10000
participants_category

Count of protests by protester demand



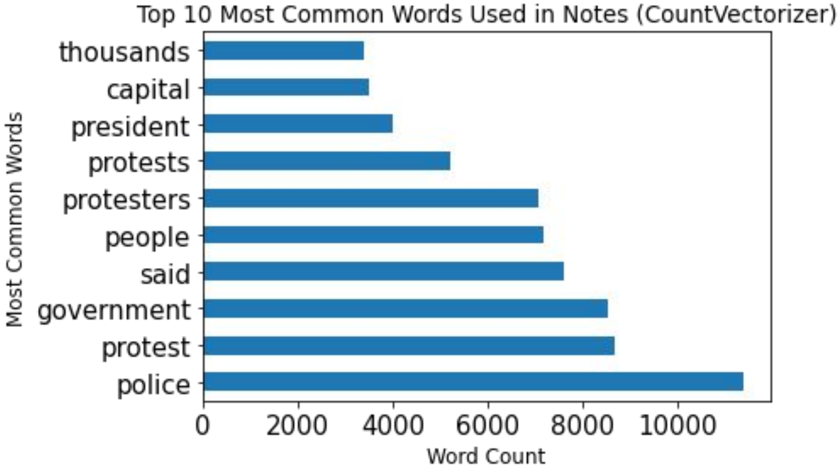
EDA - NLP

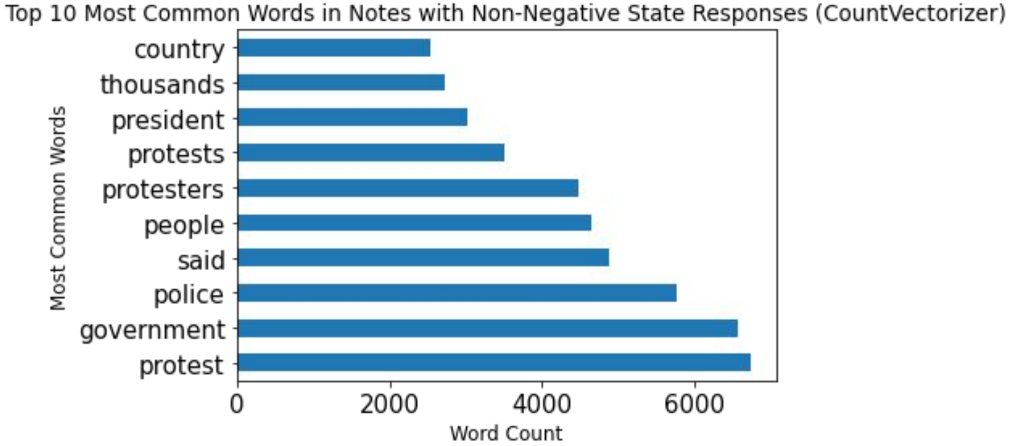

Word Count per Notes
Notes
Word Count
0
200
400
600
800
1000
0
2000
4000
6000
8000
10000
12000
14000
state_response_2
non-negative
negative

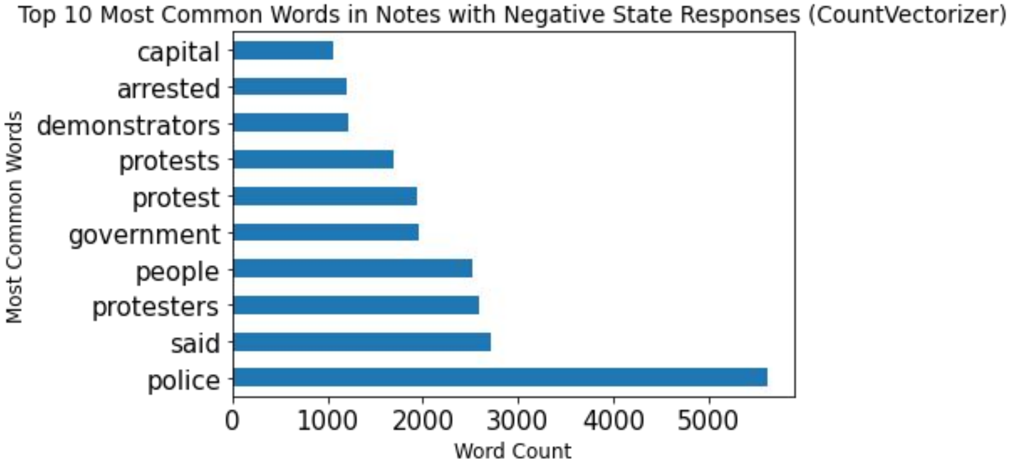

Character Count per Notes
Notes
Character Count
0
200
400
600
800
1000
1200
0
500
1000
1500
2000
state_response_2
non-negative
negative

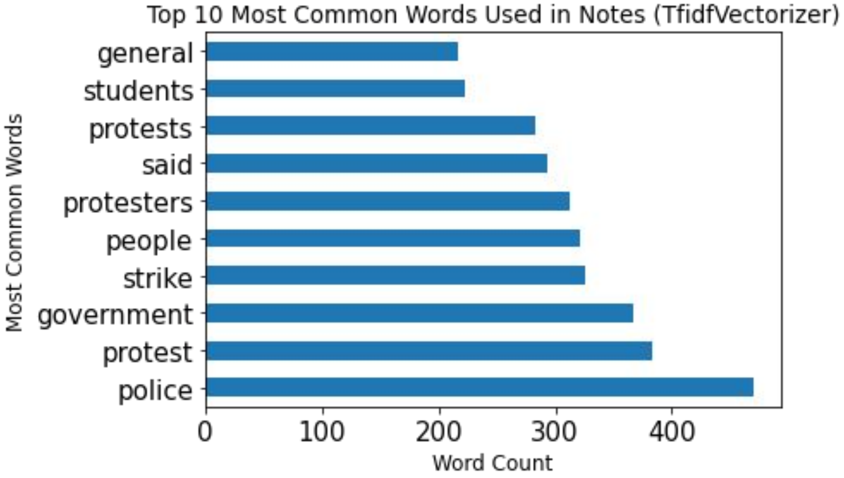
WordCloud

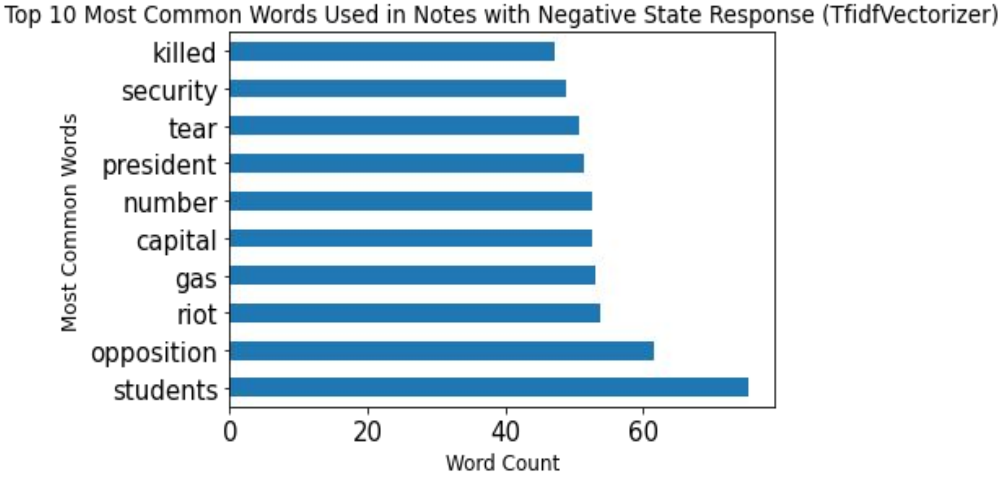

papua new
government
protest
prime minister
new guinea
canada
canadian
indian
asylum seeker
violent
violence
federal
quebec
bougainville
amid
people
service
centre
mohawk
police
will
yesterday
officials
affairs
montreal
half
minister
announced
passenger
queen
detention
leave
student
island
cut
run
clashes
manus
said
parliament
demanding
licence
peter
vote
neill
mining
fears

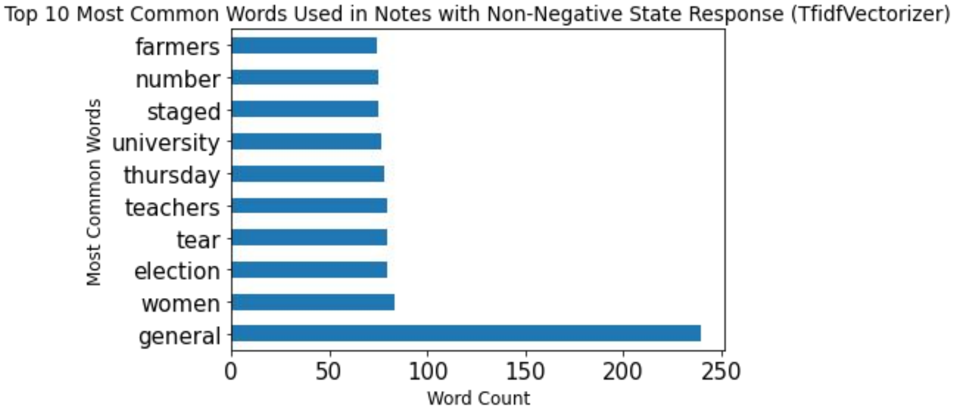

Top 10 Most Common Words Used in Notes (CountVectorizer)
Most Common Words
thousands
capital
president
protests
protesters
people
said
government
protest
police
0
2000
4000
6000
8000
10000
Word Count


Top 10 Most Common Words in Notes with Non-Negative State Responses (CountVectorizer)
Most Common Words
country
thousands
president
protests
protesters
people
said
police
government
protest
0
2000
4000
6000
Word Count


Top 10 Most Common Words in Notes with Negative State Responses (CountVectorizer)
Most Common Words
capital
arrested
demonstrators
protests
protest
government
people
protesters
said
police
0
1000
2000
3000
4000
5000
Word Count


Top 10 Most Common Words Used in Notes (TfidfVectorizer)
Most Common Words
general
students
protests
said
protesters
people
strike
government
protest
police
0
100
200
300
400
Word Count


Top 10 Most Common Words Used in Notes with Negative State Response (TfidfVectorizer)
Most Common Words
killed
security
tear
president
number
capital
gas
riot
opposition
students
0
20
40
60
Word Count


Top 10 Most Common Words Used in Notes with Non-Negative State Response (TfidfVectorizer)
Most Common Words
farmers
number
staged
university
thursday
teachers
tear
election
women
general
0
50
100
150
200
250
Word Count

Modeling - Categorical Features (2 & 3 Classes)

- The modeling results were categorized in 4 ways:
  - 2 classes using year data
  - 3 classes using year data
  - 2 classes without year data
  - 3 classes without year data
- Class imbalance techniques were tested for each category:
  - Oversampling the least frequent class
  - Undersampling the most frequent class
  - Weighted models
- Optimizing for accuracy, but precision taken into account

Model Insights - Categorical Features (2 & 3 Classes)

- Baseline Models
  - Weighted models performed best: Logistic Regression, Support Vector Classifier & XGBoost
  - 2 classes performed best with accuracy (mid-high 70's), 3 classes with precision (low 60's)
  - Year data made no difference
- Tuned Models
  - None of the models performed much better than the baseline models, some performed worse
  - Overall best-performing model: Support Vector Classifier (2 classes) → Accuracy: 0.771 | Variance: 0.003
  - Year data made no difference

Model Insights - Categorical Features (6 Classes)

	Train Accuracy	Test Accuracy
XGBoost	0.54	0.42
Neural Networks	0.43	0.41

- Comparable test performance
- Less variance for neural networks

Model Insights - NLP

- Logistic Regression
  - Best performer & most efficient
  - 82% accuracy
  - 83% precision
- XGBClassifier
  - Just as good as Logistic Regression but less efficient
  - 81% accuracy
  - 83% precision
- Multinomial NB
  - 75% accuracy
  - 83% precision

Other Findings:

- Oversampler worked best than undersampling on all models worked on

Conclusion

- NLP based model performed better than using the categorical features predictors
- Logistic regression is the best performer with highest efficiency

Recommendations & Next Steps

Detailed open source text analysis

Numerical features