Classifying Protests by State Response

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?

Dataset

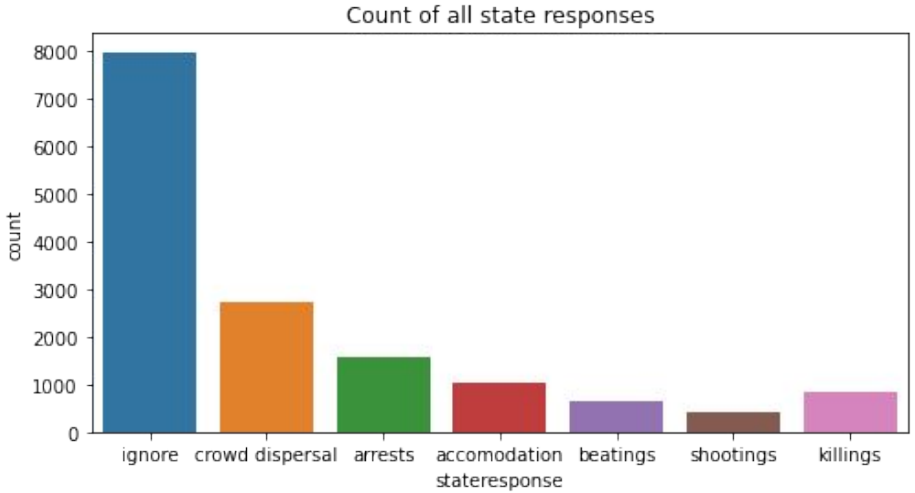
Mass Mobilization Protest Data *from Harvard Dataverse*

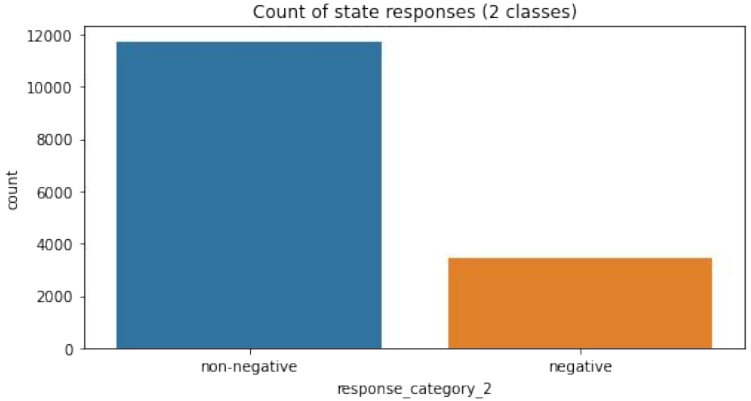
- Protests from 162 countries between 1990 and March 2020
- 15198 instances of mass mobilization events

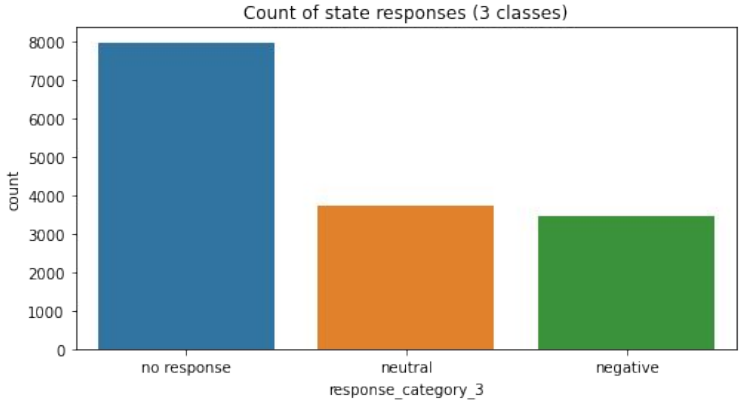
Data Cleaning - Categorical Features

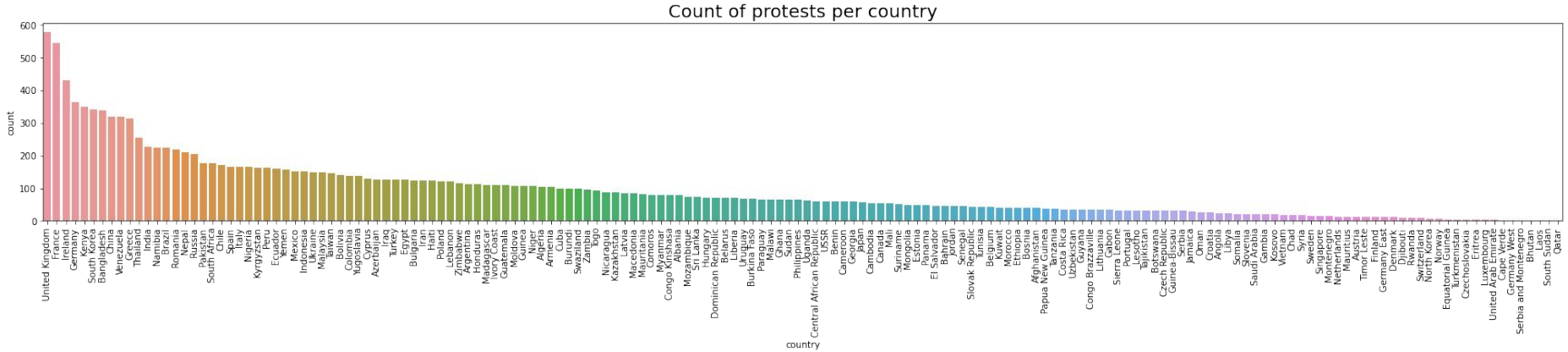
- The original dataset consisted of mainly categorical features
- Features focused on during cleaning:
  - ‘protest’
  - ‘startday’, ‘startmonth’, ‘startyear’, ‘endday’, ‘endmonth’ and ‘endyear’
  - ‘participants_category’ and ‘participants’
  - ‘stateresponse1’ - stateresponse7’
  - ‘protesterdemand1’ - protesterdemand4’

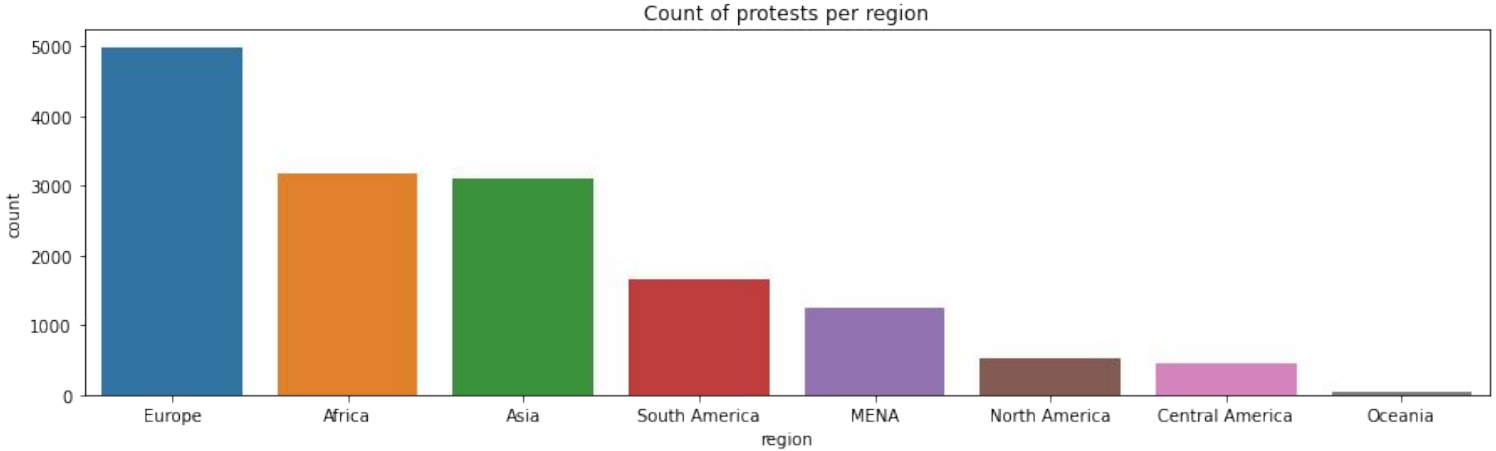
EDA - Categorical Features

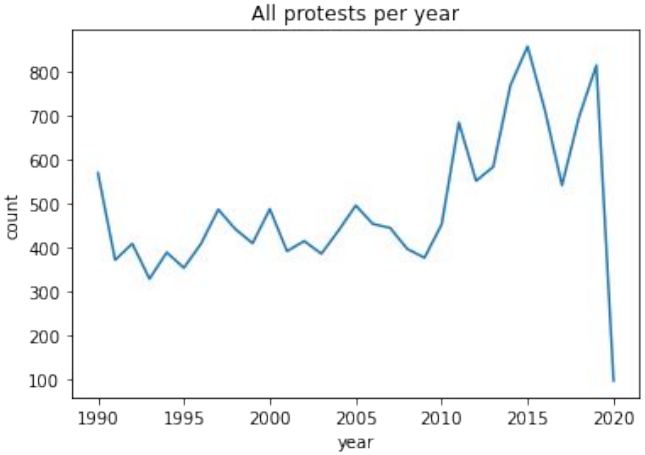


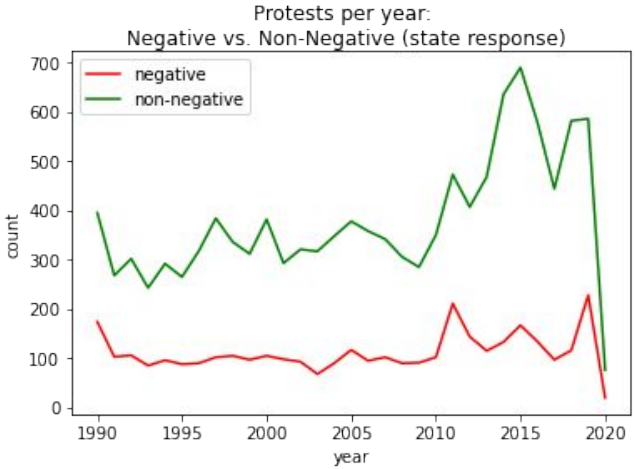


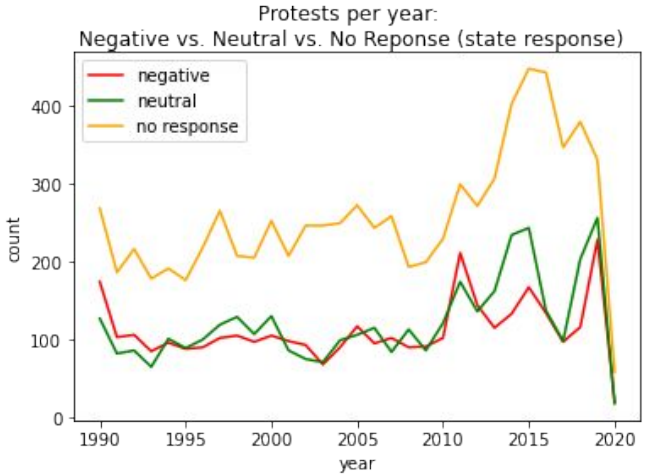


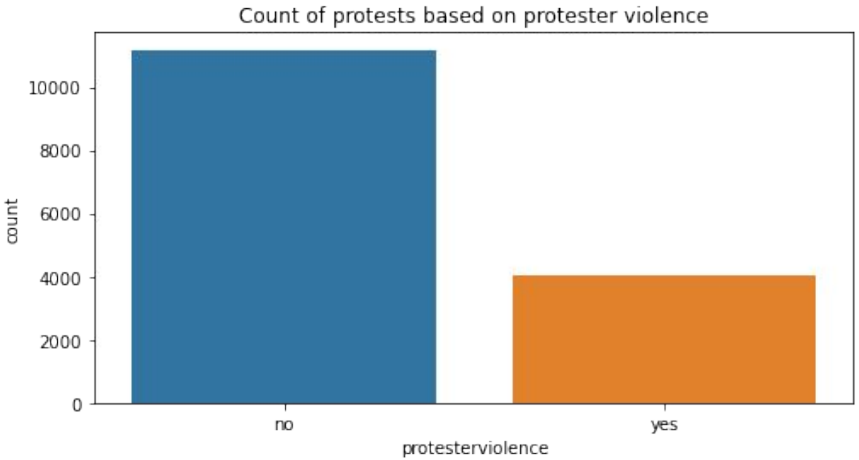

Count of protests per country
count
country

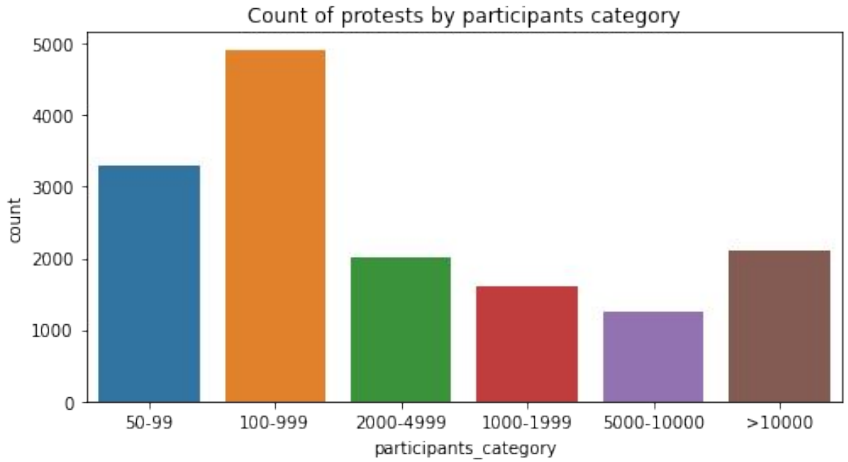

Count of protests per region
count
region
Europe
Africa
Asia
South America
MENA
North America
Central America
Oceania


All protests per year
count
year
100
200
300
400
500
600
700
800
1990
1995
2000
2005
2010
2015
2020

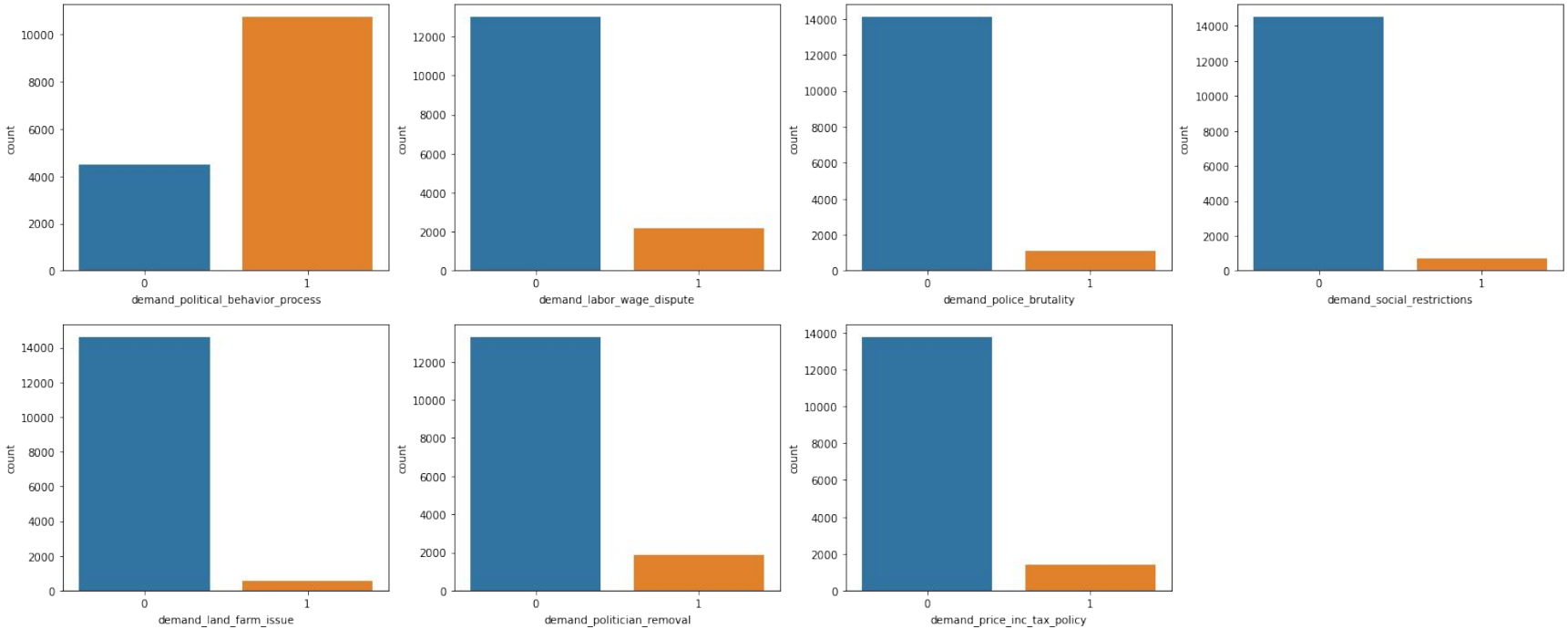

Protests per year:
Negative vs. Non-Negative (state response)
negative
non-negative
count
year
0
100
200
300
400
500
600
700
1990
1995
2000
2005
2010
2015
2020


Protests per year:
Negative vs. Neutral vs. No Reponse (state response)
negative
neutral
no response
count
year
0
100
200
300
400
1990
1995
2000
2005
2010
2015
2020

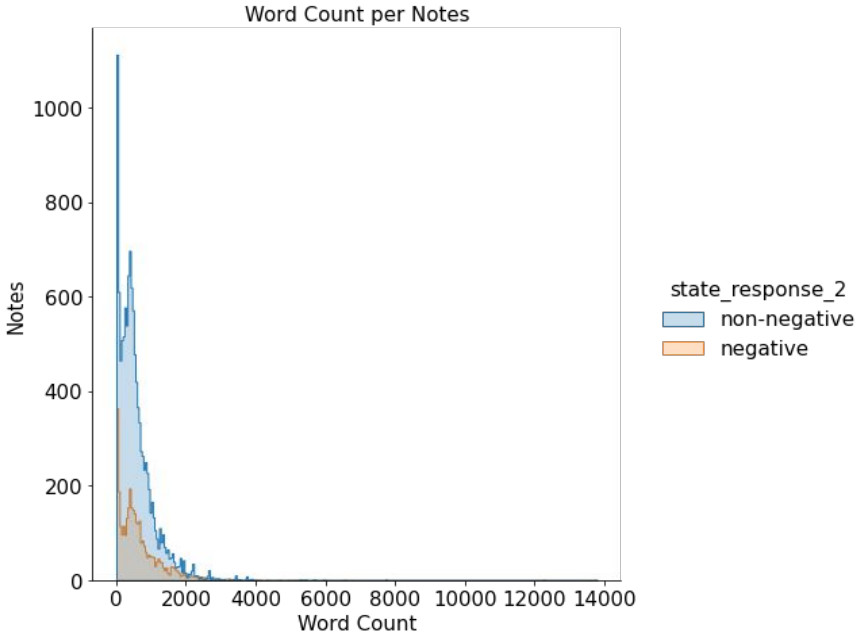

Count of protests based on protester violence
count
protesterviolence
no
yes
0
2000
4000
6000
8000
10000

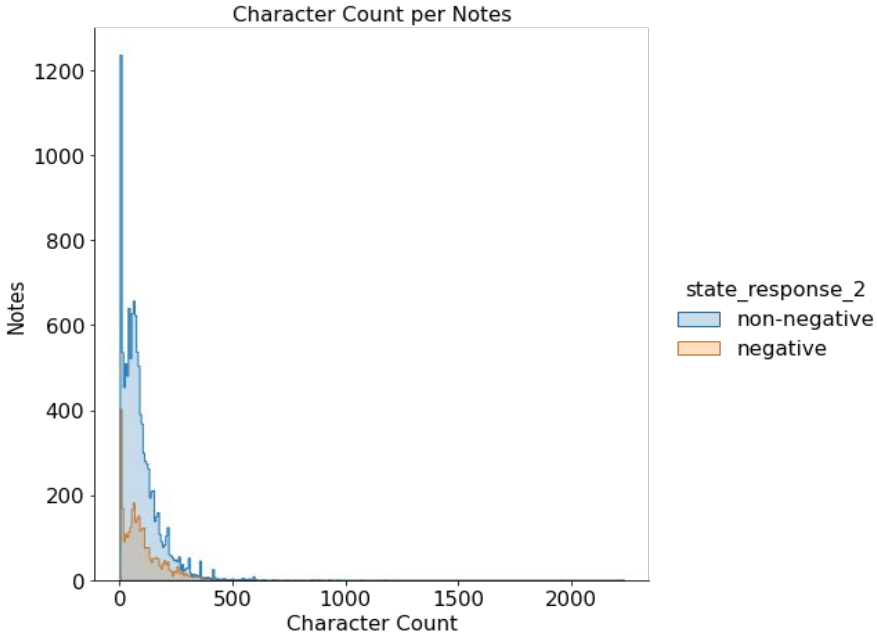

Count of protests by participants category
count
participants_category
50-99
100-999
2000-4999
1000-1999
5000-10000
>10000
0
1000
2000
3000
4000
5000

Count of protests by protester demand



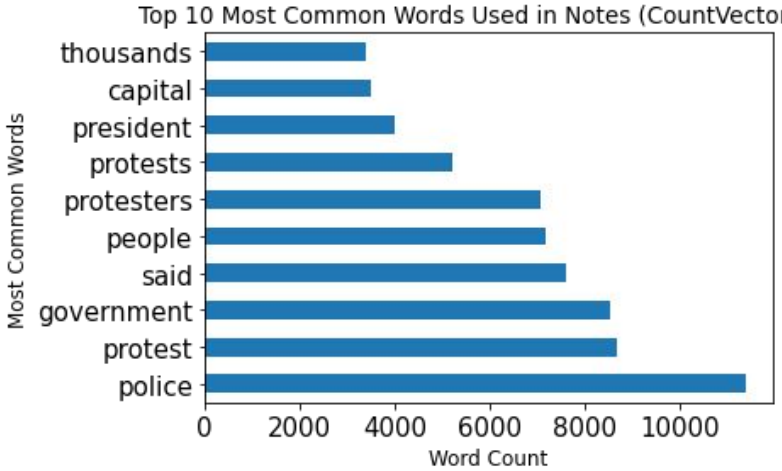
EDA - NLP

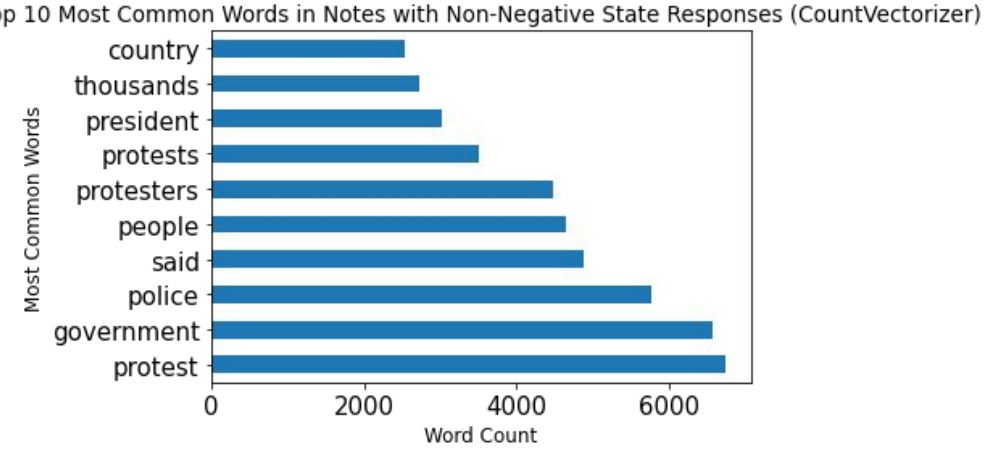

Word Count per Notes
Notes
Word Count
0
200
400
600
800
1000
0
2000
4000
6000
8000
10000
12000
14000
state_response_2
non-negative
negative

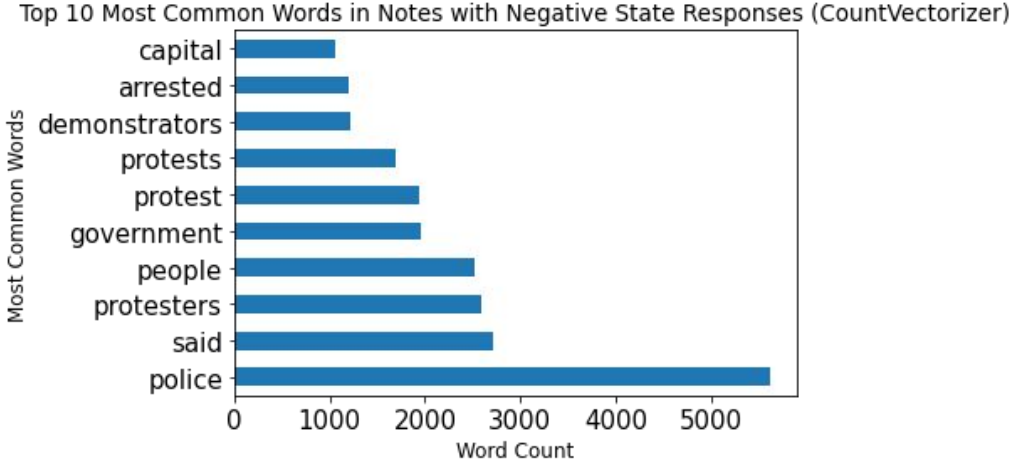

Character Count per Notes
Notes
Character Count
0
200
400
600
800
1000
1200
0
500
1000
1500
2000
state_response_2
non-negative
negative

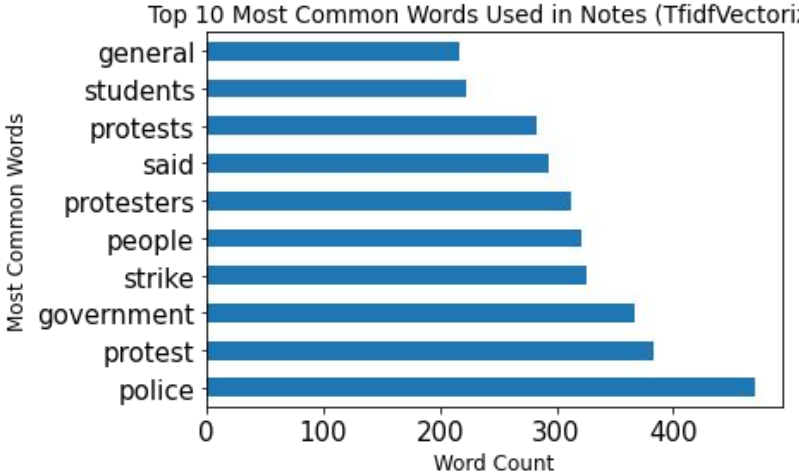
WordCloud

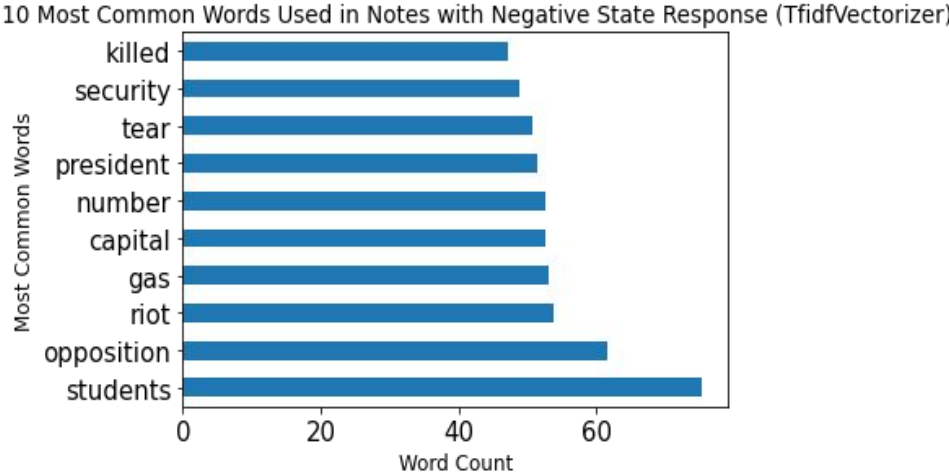

papua new
protest
government
federal
service
amid
bougainville
centre
mohawk
prime minister
canada
new guinea
violent
violence
quebec
indian
canadian
yesterday
officials
queen
passenger
run
asylum seeker
people
affairs
montreal
half
minister
announced
police
will
fears
demanding
owned
finally
identified
eliminated
crisis
licence
peter
student
manus
clashes
cut
island
mining
detention
leave
said
parliament
celebrate
thousands
routes
spontaneous
vote
neill
arrive
although
railway
tomorrow
holiday
hour
now
regions
park
last
stations
protesters
made
system
province
national
urged
gathered
entered
along
approved
earlier
accompanied
braved
french
strolled
young
help
remain
without
decision
agreed
subject
calling
reports
united
terms
protestors
dispute
rail
acquire
land
mounted
entire
via
today
purchase
quebecers
gathering
constitutional

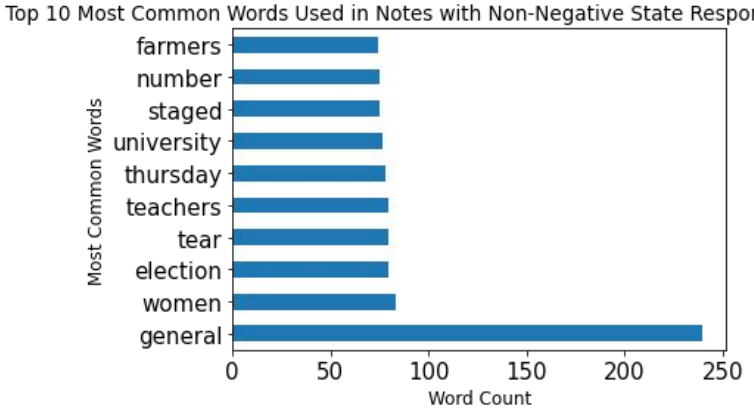

Top 10 Most Common Words Used in Notes (CountVectorizer)
thousands
capital
president
protests
protesters
people
said
government
protest
police
Most Common Words
0
2000
4000
6000
8000
10000
Word Count


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


Top 10 Most Common Words in Notes with Negative State Responses (CountVectorizer)
capital
arrested
demonstrators
protests
protest
government
people
protesters
said
police
Most Common Words
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 with Near Miss
  - 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
- Conclusion
  - All categorical feature dataset was not ideal
  - Unable to achieve a decent accuracy score

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