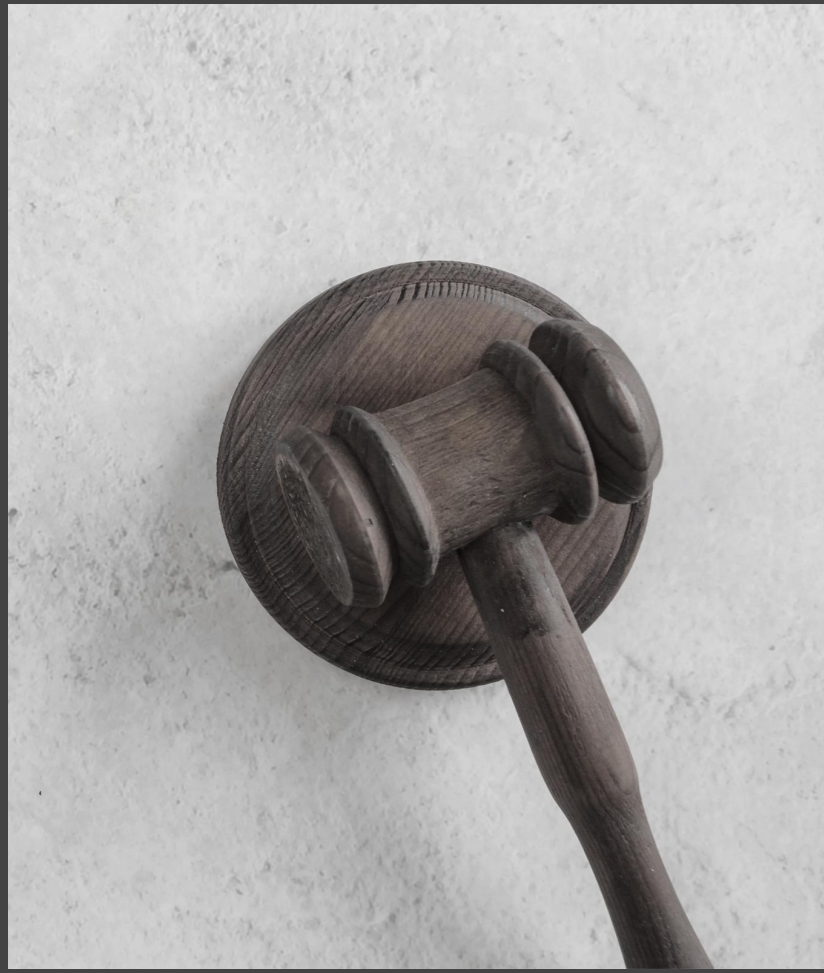
Classification of U.S. Supreme Court Opinions

Using various NLP techniques.



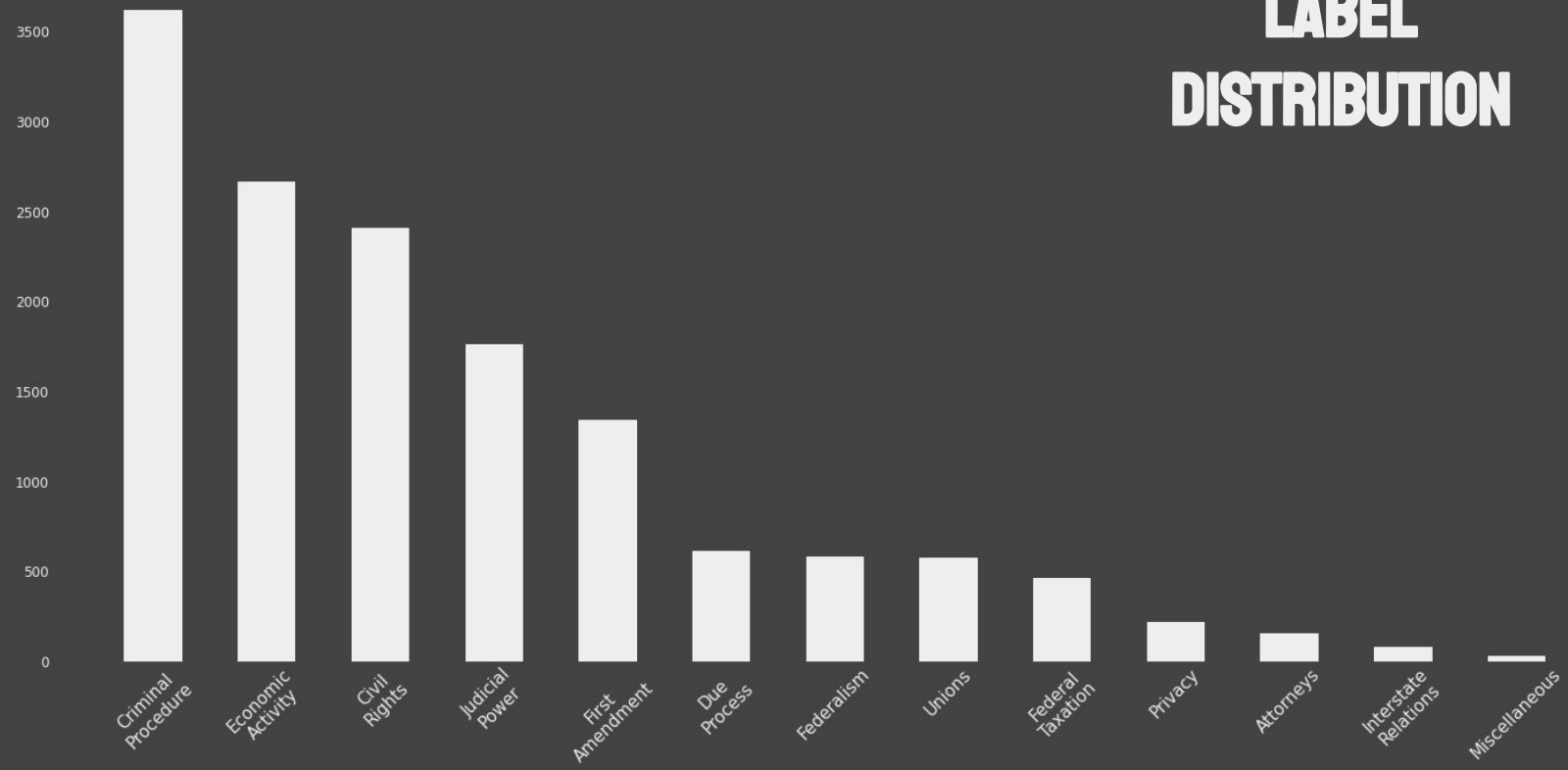
INTRODUCTION

- Thousands of court opinions are published each year.
- They have to be manually analyzed and categorized to facilitate research.
- If we can automate this process it can dramatically lower costs.

PROCESS

- Full text of some 8000 Supreme Court was gathered
- Labels were added to the opinions which classified into categories.
    - 13 Categories
- We then tried to see if various NLP Algorithms could accurately classify opinions into the right category
- We also tried to use unsupervised learning to recreate these topics.

LABEL DISTRIBUTION



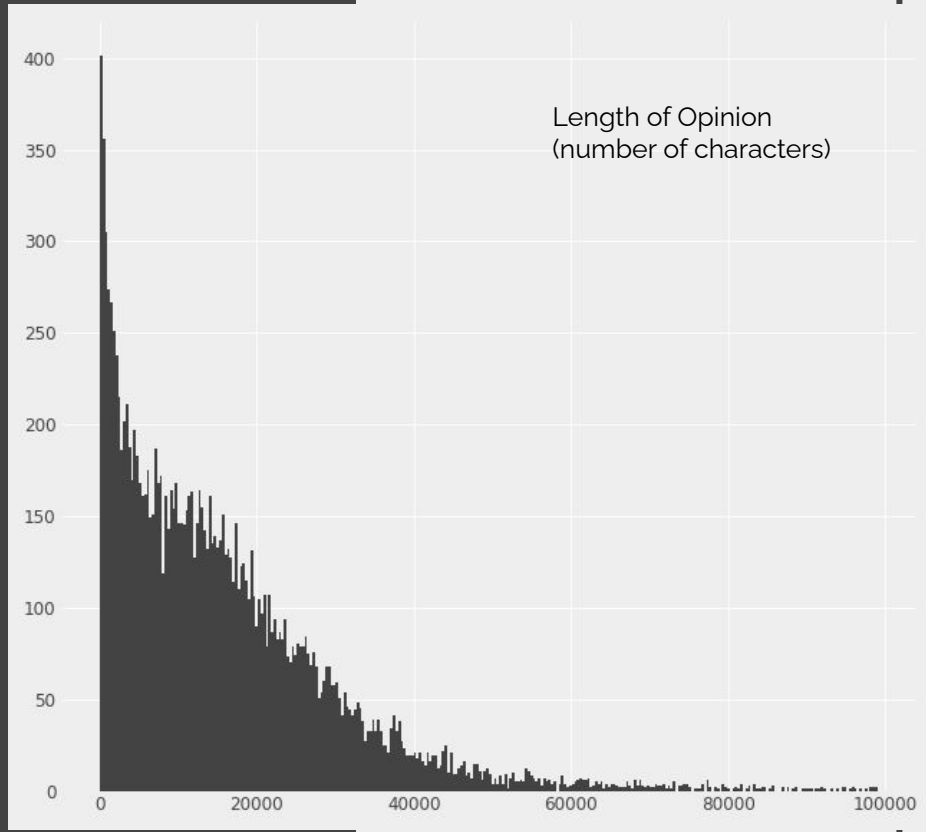
In order to minimize noise and improve model performance, we narrowed down the opinions to build our classifier.

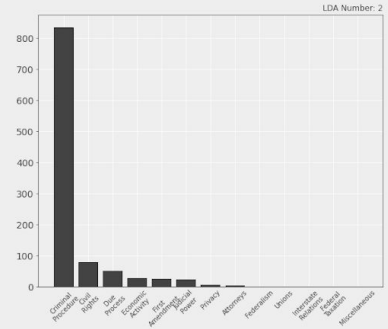
We only kept opinions whose lengths were;

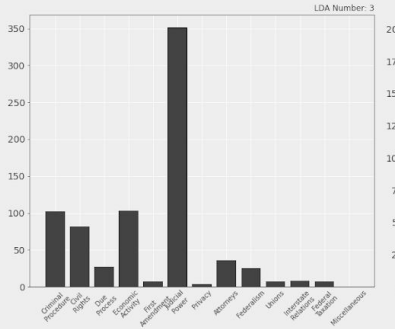
- greater than 5000 characters;
  - Alot of these are per curiam dismissals and affirmations of lower court decisions with no substance.
- less than 85000 characters;
  - This leaves out unusually long opinions

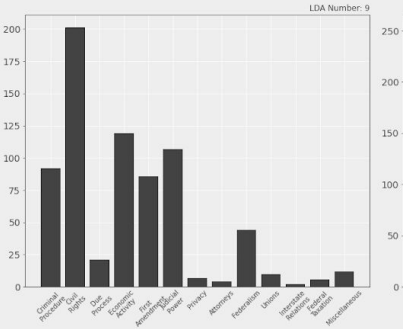
We also filtered out dissents, since they discuss the same subject matter, so seem redundant for the purposes of classification and may add noise.

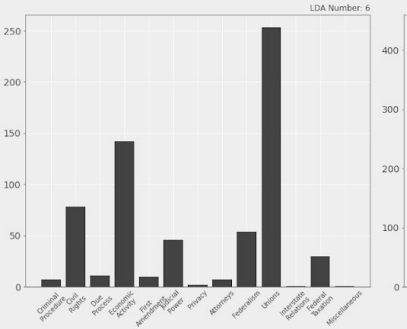
This leaves us with 6,329 opinions to use for training and testing our model.

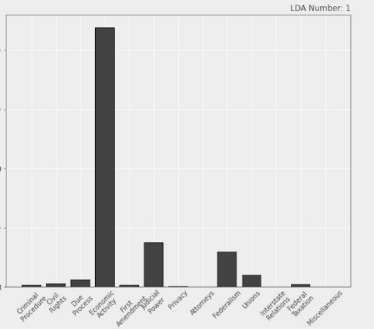


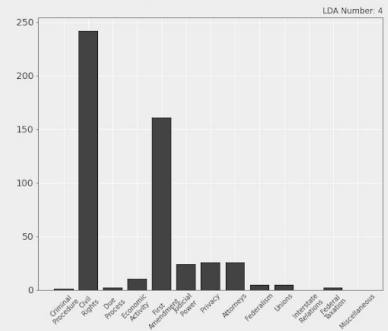


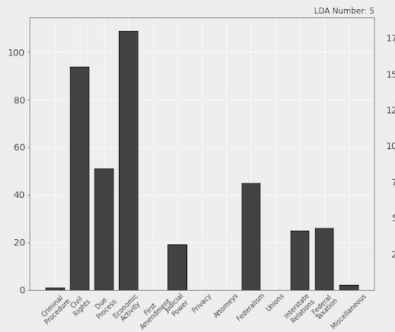


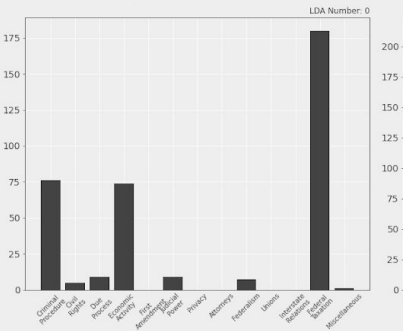


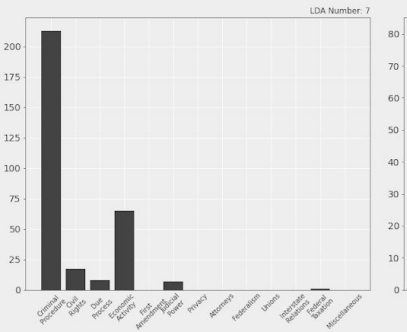


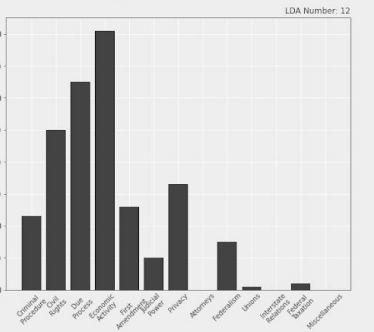


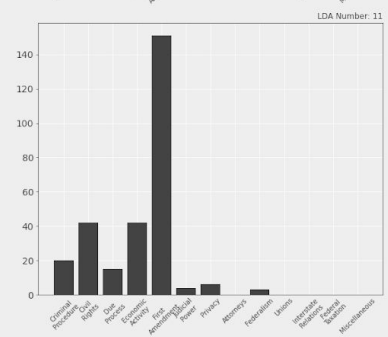


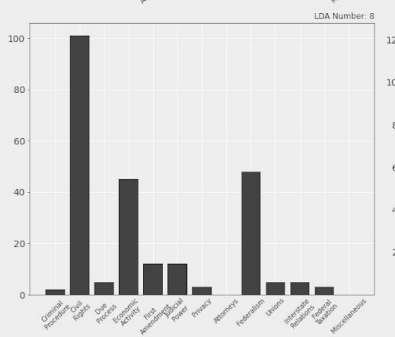


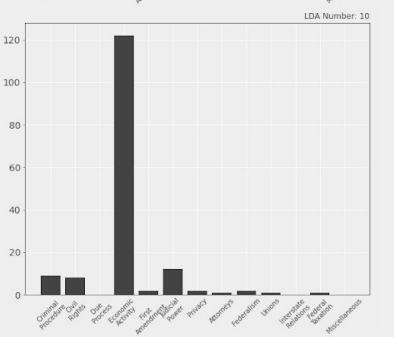




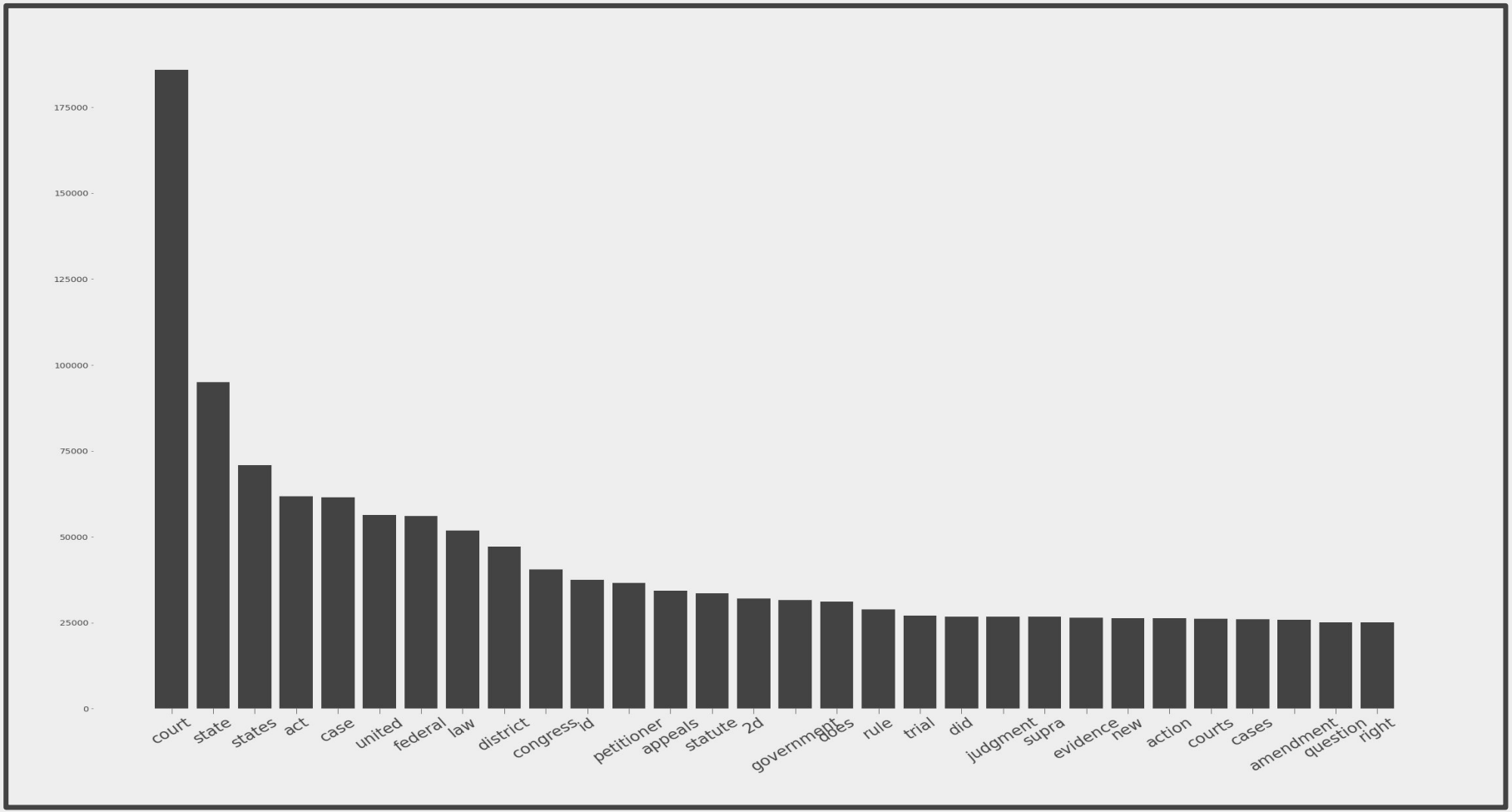








DO LDA COMPONENTS CAPTURE EXISTING LABEL CLUSTERS?


175000
150000
125000
100000
75000
50000
25000
0
court
state
states
act
case
united
federal
law
district
congress
id
petitioner
appeals
statute
2d
government
does
rule
trial
did
judgment
supra
evidence
new
action
courts
cases
amendment
question
right

CLASSIFICATION PERFORMANCE

Description	F1_score	Accuracy	Model	Vectorizer	stopwords
SVC/TFIDF and english st	0.801	0.806	LinearSVC	TfidfVectorizer	english
Sup Vec with TFIDF	0.797	0.804	LinearSVC	TfidfVectorizer	
SVC/TFIDF combined sto	0.795	0.801	LinearSVC	TfidfVectorizer	English + Most Comm
SVC/TFIDF and common	0.794	0.802	LinearSVC	TfidfVectorizer	Most Common Words
Base SupVector	0.770	0.773	LinearSVC	CountVectorizer	
Base MNB	0.757	0.766	MultinomialNB	CountVectorizer	
Base Random Forest	0.615	0.665	RandomForestClassifier	CountVectorizer	
base KNN	0.610	0.661	RandomForestClassifier	CountVectorizer	
Glove	0.589	0.404	LinearSVC	customW2V class	
Neural Net	0.580		NeuralNet	padded_sequence	

RECOMMENDATIONS & FUTURE WORK

More Training Data

- cases from other courts and jurisdictions

Use advanced pre-trained models

- BERT

Predict other targets.

- Judges Ideological leanings and voting tendencies.

THANKS!

Do you have any questions?

umarkhan314@outlook.com
texasanalytics.net