Classification of U.S. Supreme Court Opinions

Using various NLP techniques.

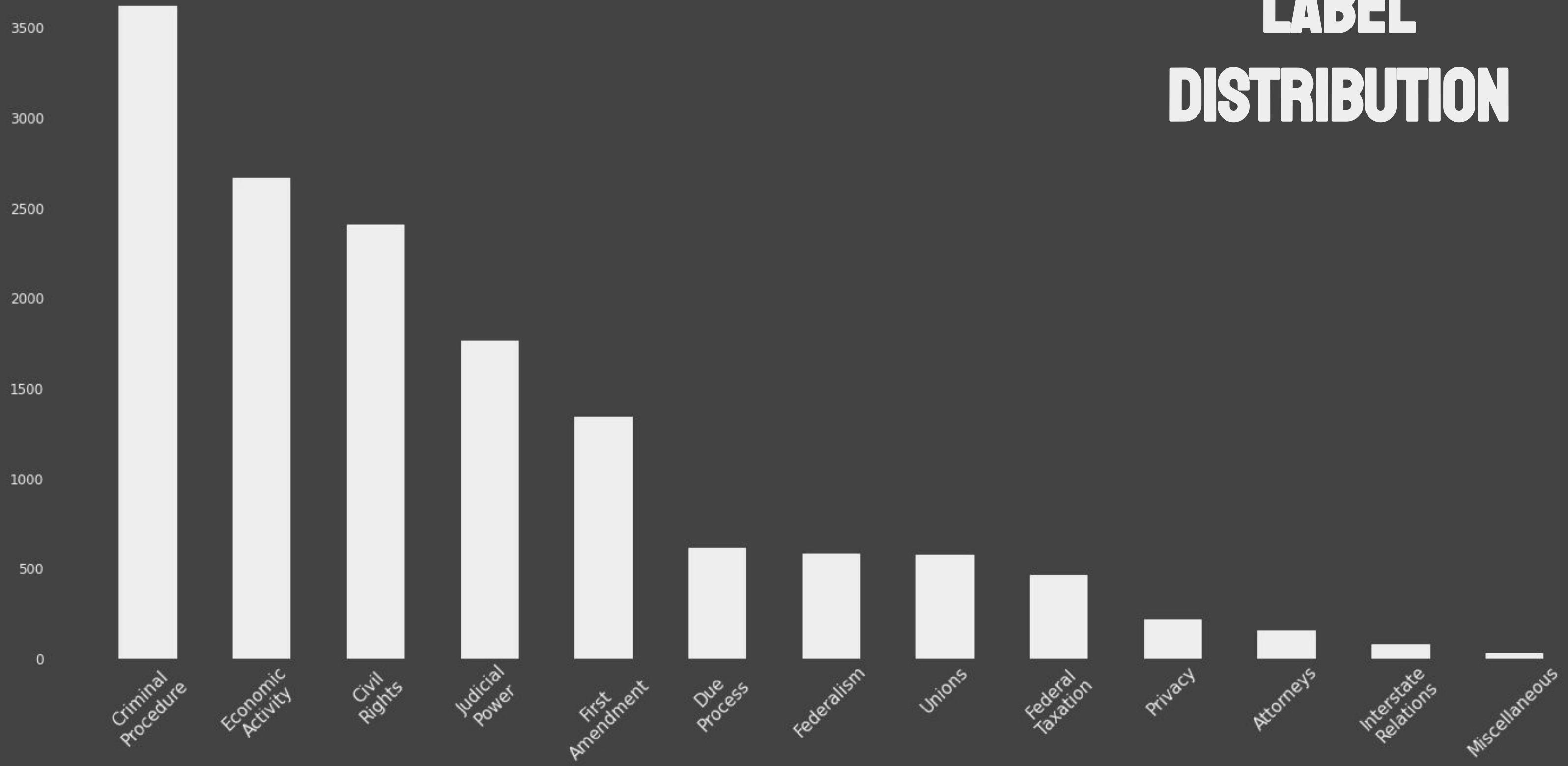
INTRODUCTION

- 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



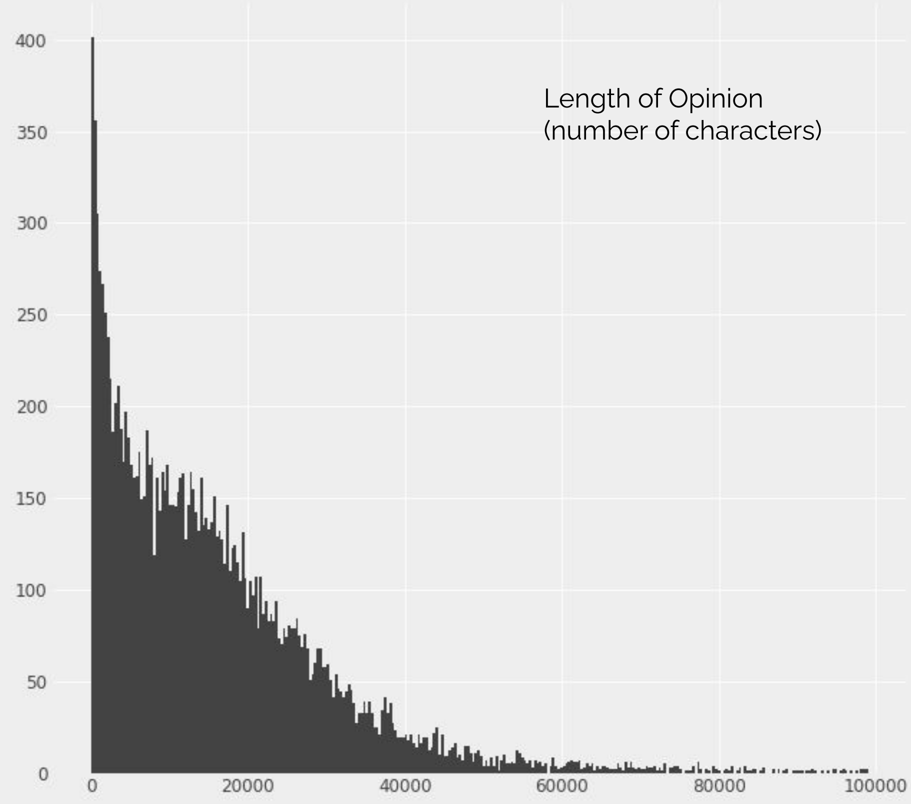
In order to minimize noise and improve model performance, we narrowed down the opinions to build our classifier.

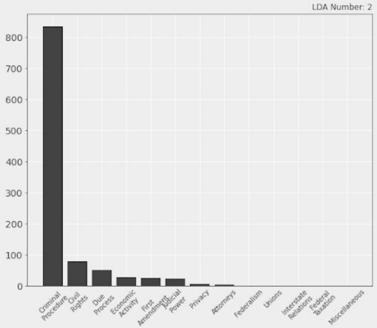
We only kept opinions whose lengths were;

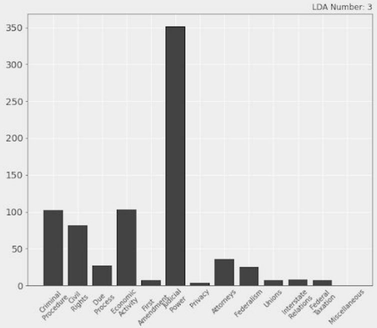
- greater than 5000 characters;
  - Alot of these are per curiam dismissals and affirmations of lower court decisions with no substance.
- less than 85000 characters;
  - This leaves out unusually long opinions

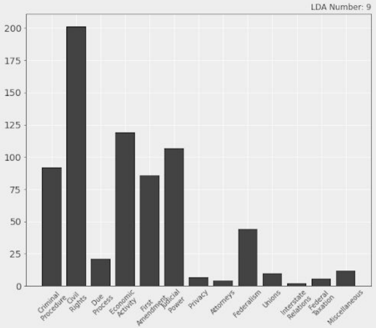
We also filtered out dissents, since they discuss the same subject matter, so seem redundant for the purposes of classification and may add noise.

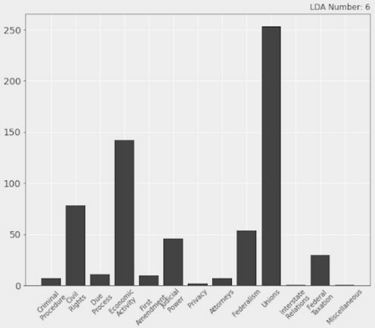
This leaves us with 6,329 opinions to use for training and testing our model.

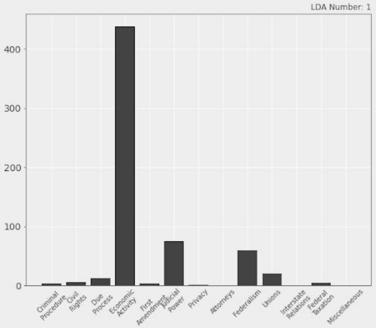


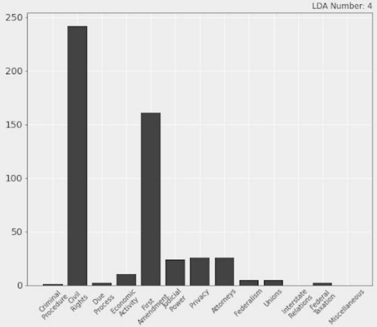


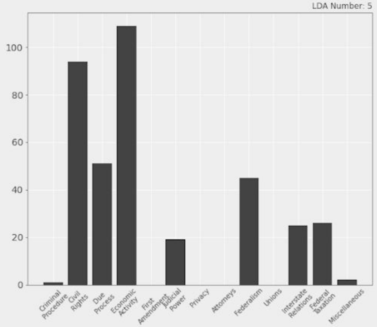


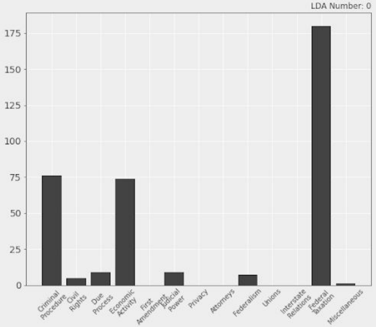


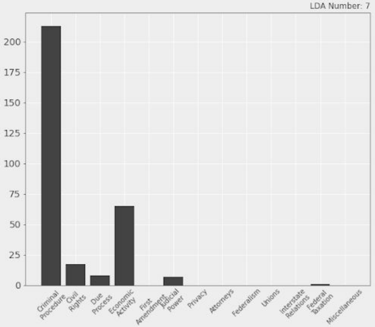


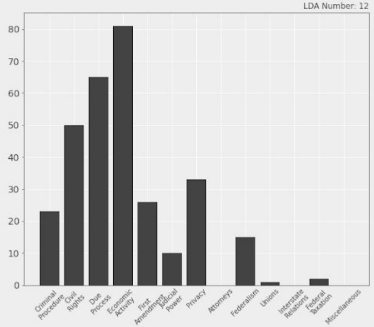


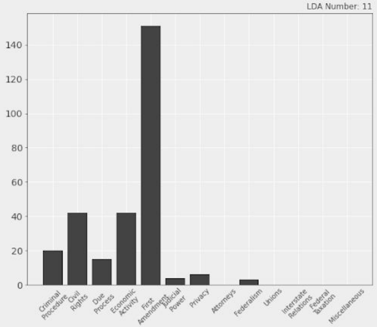


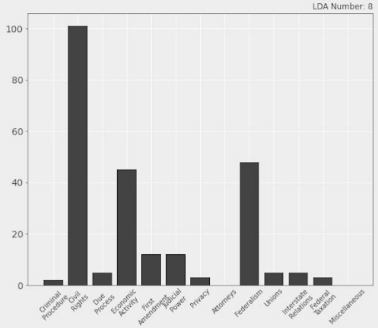


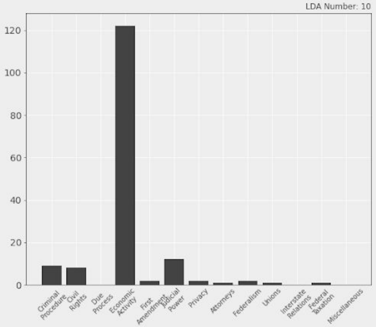




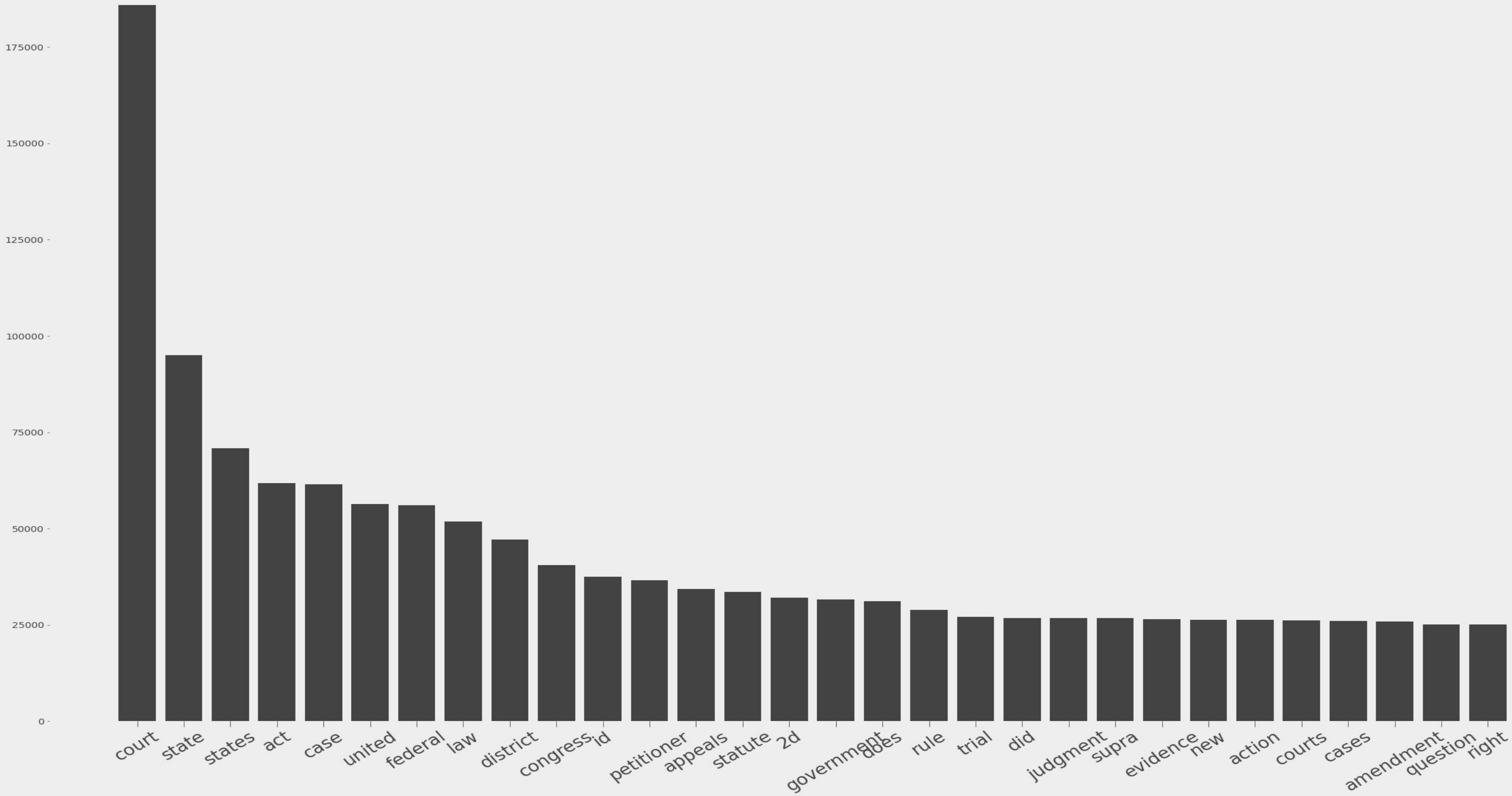




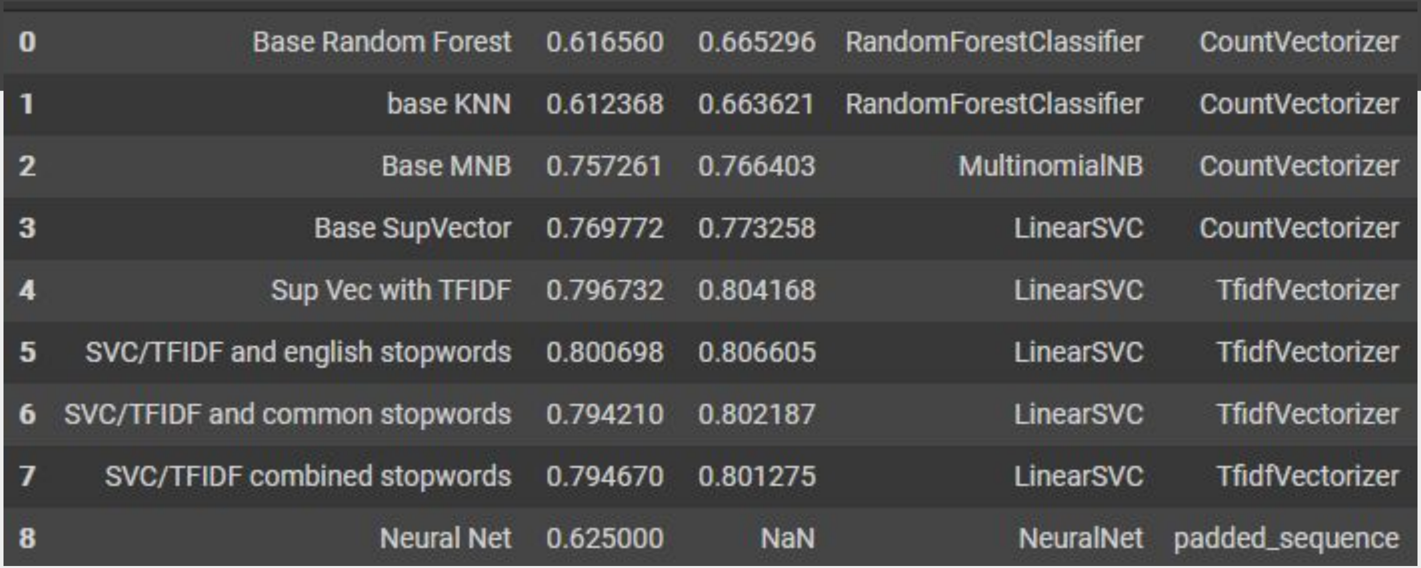




DO LDA COMPONENTS CAPTURE EXISTING LABEL CLUSTERS?


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



0	Base Random Forest	0.616560	0.665296	RandomForestClassifier	CountVectorizer
1	base KNN	0.612368	0.663621	RandomForestClassifier	CountVectorizer
2	Base MNB	0.757261	0.766403	MultinomialNB	CountVectorizer
3	Base SupVector	0.769772	0.773258	LinearSVC	CountVectorizer
4	Sup Vec with TFIDF	0.796732	0.804168	LinearSVC	TfidfVectorizer
5	SVC/TFIDF and english stopwords	0.800698	0.806605	LinearSVC	TfidfVectorizer
6	SVC/TFIDF and common stopwords	0.794210	0.802187	LinearSVC	TfidfVectorizer
7	SVC/TFIDF combined stopwords	0.794670	0.801275	LinearSVC	TfidfVectorizer
8	Neural Net	0.625000	NaN	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

CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, and infographics & images by Freepik.