Politics On Reddit: A Journey into the American Political Landscape

OH NO!

AmericanPoliticsDebate.com has been hacked! Hackers got into our system and mixed up our Conservative and Liberal discussion groups. Now, we need to create a model that can classify discussions as Liberal or Conservative and get our system back in order.

01
Data Gathering
Using the Reddit API

02
Data Cleaning
Removing Unwanted Data

03
EDA
Describe the topic of the
Exploring the Data

04
Modeling
Fitting and Scoring
Models

05
Modeling
Evaluation
Understanding
Misclassification

4. We are not *fair and balanced*. We don't pretend to be unbiased. We don't pretend to give all commenters equal time. This is *by conservatives and for conservatives*. We are here to discuss conservative topics from a distinctly conservative point of view. If you don't like that it's not an unbiased forum, go ask why /r/politics is a leftist totalitarian state. Leftists and moderates have never been welcomed here. If you wander in here and spout nonsense or insult us, don't be surprised when we ban you almost instantly.

R/CONSERVATIVE RULES

1. Civility ⌃

Rule 1: Be civil, follow any flair guidelines.

2. Racism ⌃

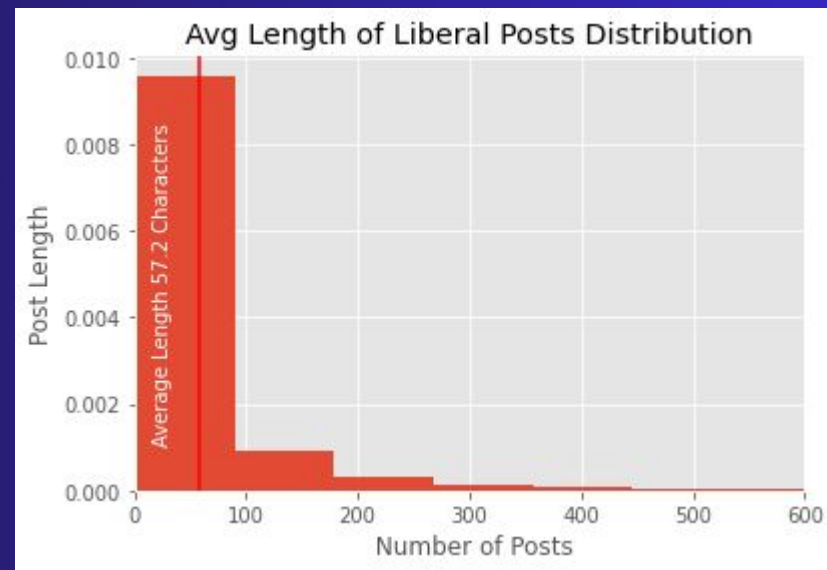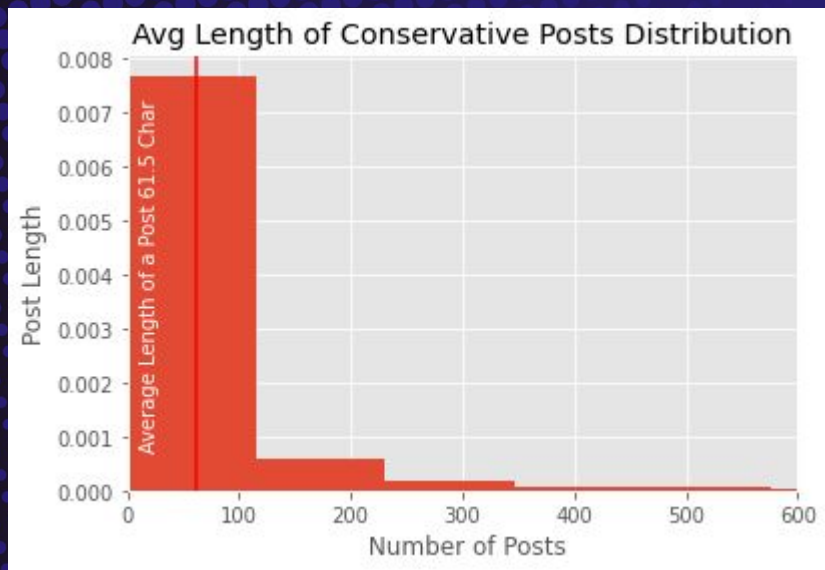Rule 2: No racism.

3. Spam ⌃

Rule 3: Don't spam article submissions (this doesn't apply to comments). Limit 4 per hour.
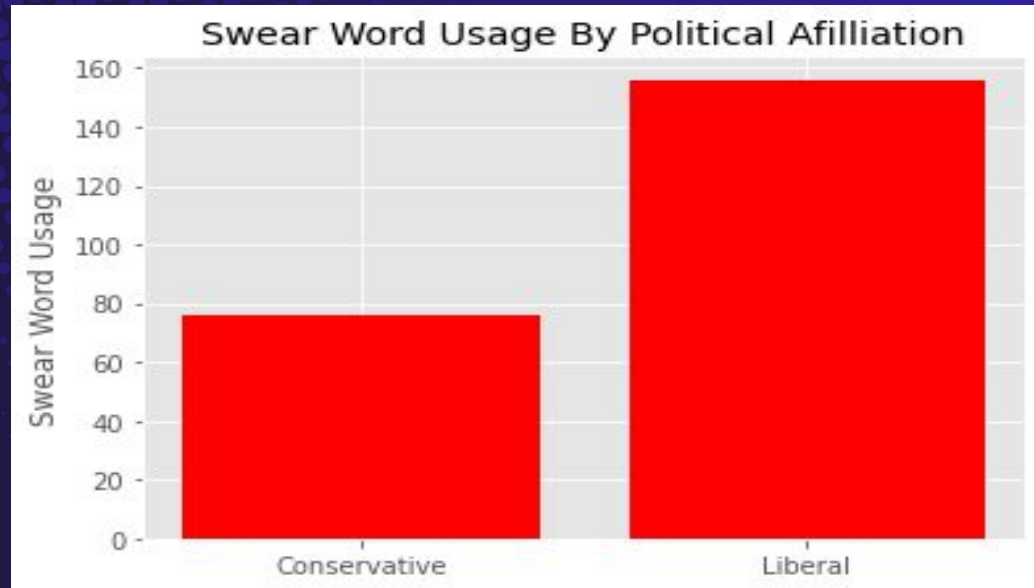
---

7. R7: No Hate Speech or Bigotry ⌃

Hate speech directed towards groups of people based on race, gender, sexual orientation, gender identity, ethnicity, religion, nationality, social status, or disability is not allowed on the sub

- Do not submit links from offensive websites

- Do not submit links from racist websites

- Do not submit links promoting hate speech

- Do not submit links from websites promoting pseudoscience (example: eugenics)

> *Note: Users that break this rule are subject to an immediate ban without a warning*

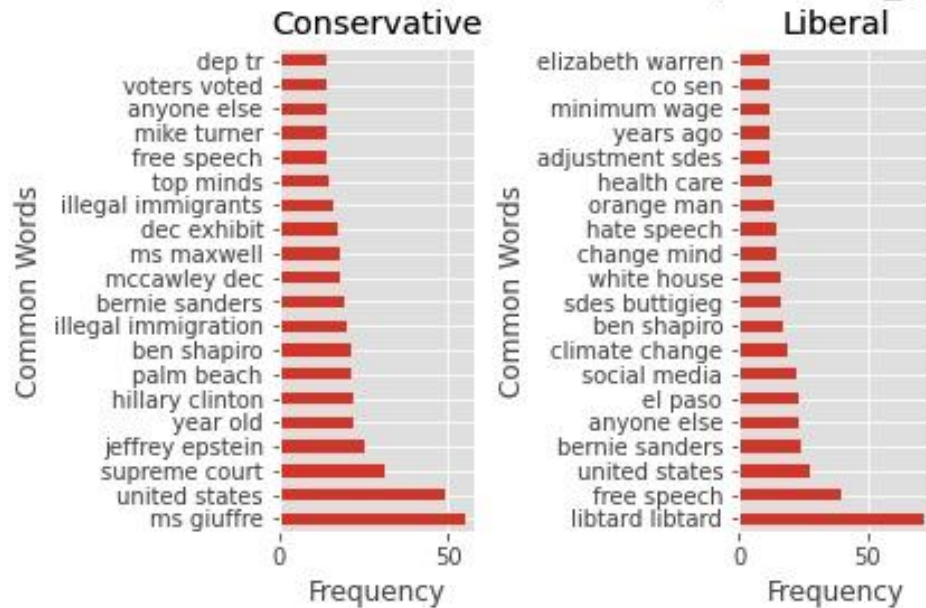**Avg Length of Conservative Posts Distribution**

Post Length / Number of Posts

Average Length of a Post 61.5 Char

Avg Length of Liberal Posts Distribution

Post Length / Number of Posts

Average Length 57.2 Characters

Swear Word Usage By Political Afilliation

Most Common Words for Each Subreddit With N_gram (1,2)

Conservative

Common Words	Frequency
first	
go	
many	
anyone	
question	
said	
going	
really	
years	
could	
obama	
make	
post	
ms	
page	
states	
need	
also	
giuffre	
epstein	

Liberal

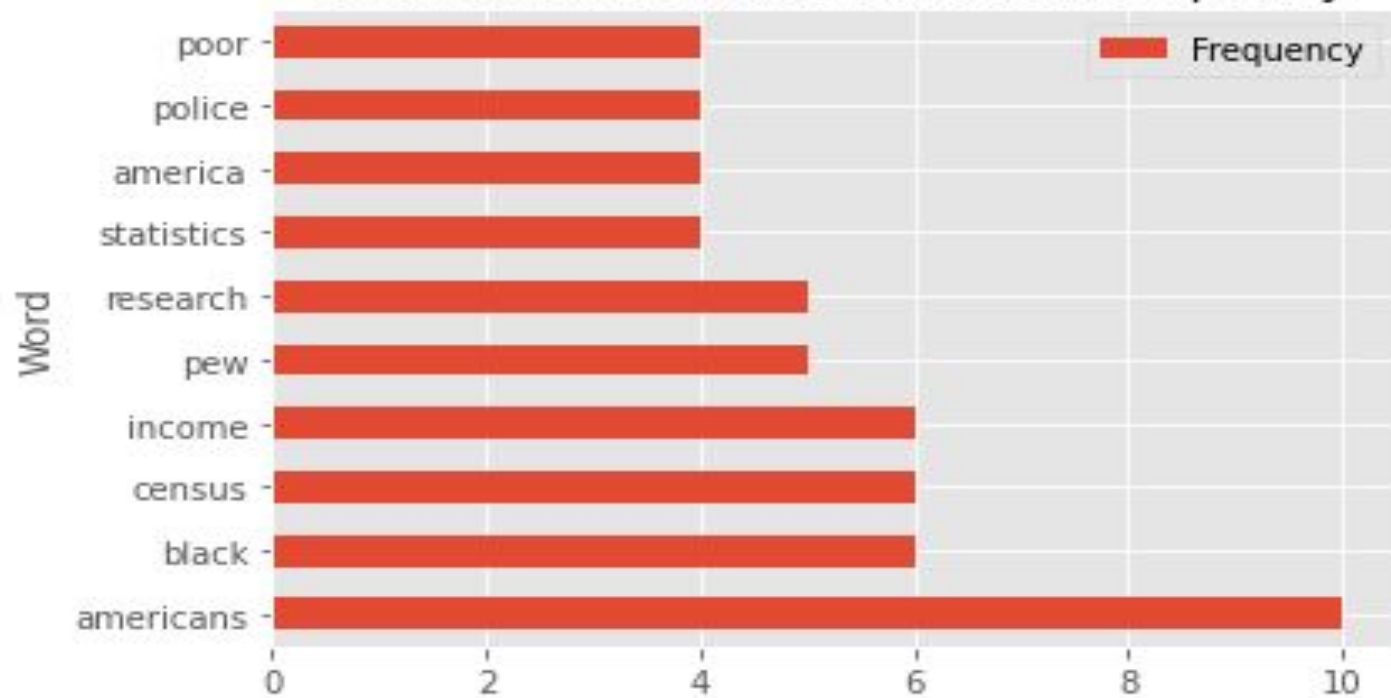| Common Words	Frequency
said	
anyone	
every	
could	
someone	
libtard	
democrats	
america	
state	
free	
help	
many	
going	
political	
need	
also	
white	
fuck	
really	
make	

20 Common Words for Each Subreddit With Stopwords N_gram(2,2)

Posted In Both Subreddits Word Frequency

Average Number of Comments By Subreddit
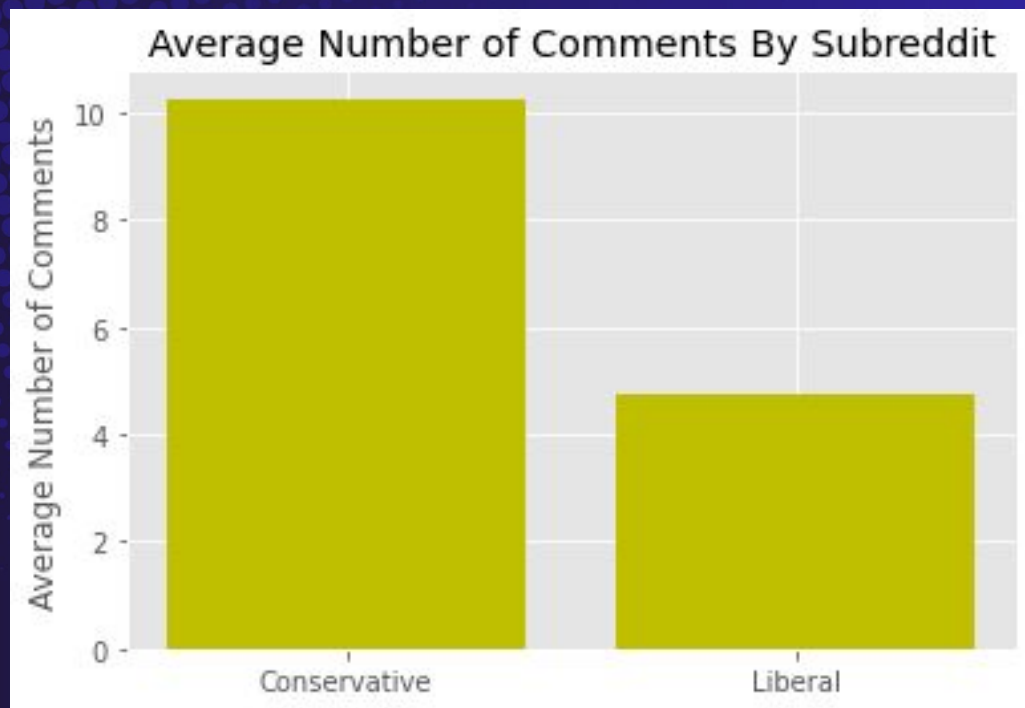
Baseline	55 - Liberal	.44 - Conservative
	Train Accuracy	Test Accuracy
Log_Regression-cvec	.88	.63
Log_Regression-tvec	.88	.665
KNN - Cvec	.63	.56
KNN - tvec	.63	.55
Decision Trees - cvec	.64	.57
Decision Trees - tvec	.65	.57
Bagging - cvec	.96	.613
Bagging - tvec	.96	.62
Random Forest cvec	.97	.638
Random Forest tvec	.97	.627
AdaBoost	.68	.61

Words Most Correlated to Liberal

Words	
charisma	
freedom	
discuss	
flag	
love trump	
single	
qatar	
virus	
women	
helped	
term	
stand	
members	
style	
liberals	
friendly	
great	
liberalism	
libtard	
libtards	
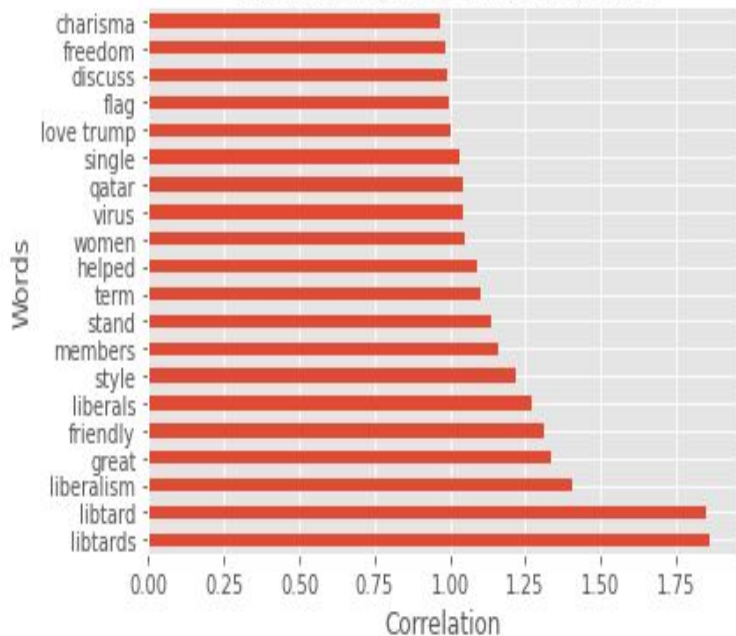
Words Most Correlated to Conservative

Words	
cruz	
deal	
deleted	
conservatives	
movement	
known	
control	
conservatism	
beat	
terms	
policies	
david	
citizen	
suddenly	
transgender	
race	
sense	
won	
investigation	
white people	
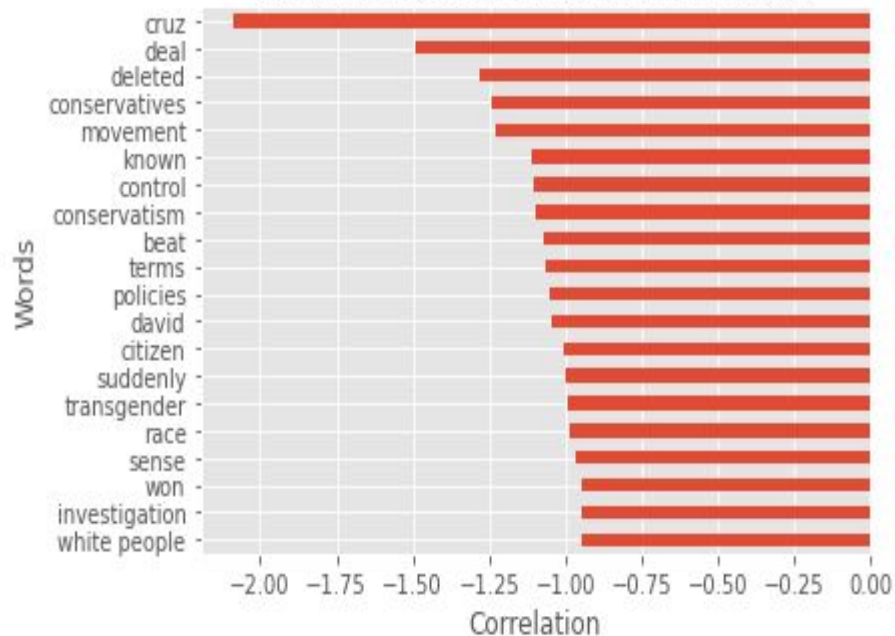
Success Metrics

	Pred Conservative	Pred Liberal
Actual Conservative	167	185
Actual Liberal	74	357

```
There are 185 False Positives
There are 167 True Negatives
There are 74 False Negatives
There are 357 True Positives
```
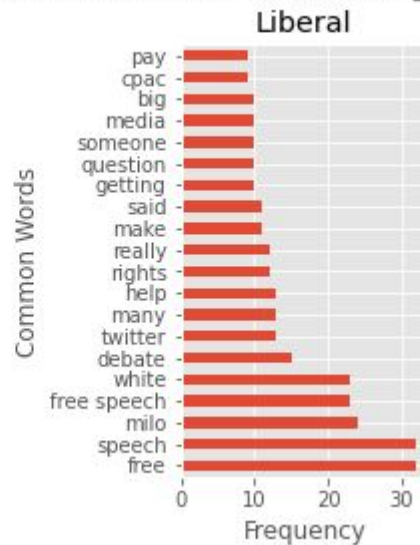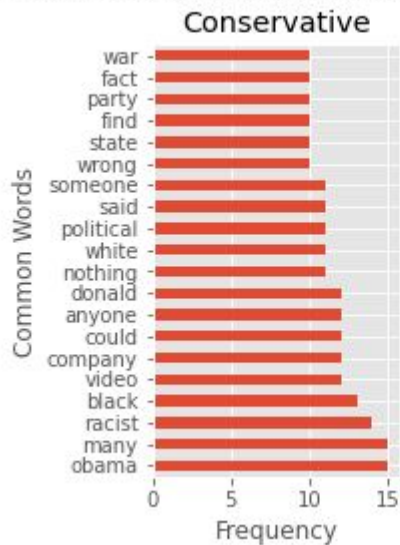
```
The Accuracy    is 0.669220945083014
The Miscassified score  is 0.330779054916986
The Precision  is 0.6586715867158671
The Sensitivity is 0.8283062645011601
The Specificity is 0.4744318181818182
```

Most Common Misclassified Posts for Each Subreddit With N_gram (2,2)
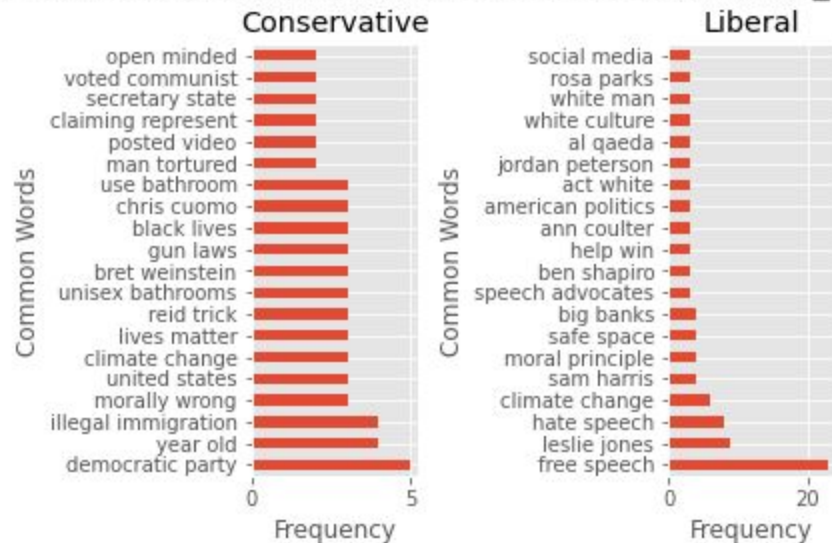
Most Common Misclassified Words for Each Subreddit With N_gram (1,2)

1. For every occurance of the word "libtards" a post is 593.0 % as likely to be in the Liberal subreddit
2. For every occurance of the words "liberals" a post is 1218.2 % as likely to be in the Liberal subreddit

1. For every occurance of the word "deleted" a post is 495.3 % as likely to be in the Conservative subreddit
2. For every occurance of the words "cruz" a post is 495.3 % as likely to be in the Conservative subreddit

Conclusions

Model Thrown Off By Conservative Trolling

01

Some of the most common words in the Liberal subreddit were attack threw off model.

Each Spend a Majority of time Talking About the Other

02

Model was thrown off by topics that were about the other group

Some issues were discussed only by one side and not by the other at all

03

For example Jeffrey Epstein Most popular in Conservative

Hard time with Large Concepts

04

Most of the misclassified posts revolved around a polarizing figure usually from the other side, for example Jordan Peterson, sam harris