Utilizing Natural Language Processing to Predict Political Affiliation

Jon Reynolds

Problem Statement

What patterns exist in online text that we can extract, analyze and use to refine online business models?

- NLP-based classification models that can predict whether the author is more likely to hold conservative or liberal views

- Inform online advertisers which ads to run for which audiences.

- r/Democrats has less subscribers with 91k as compared to /r/Conservative's 205k, but how does their word choice stack up?

What is Reddit and How Can We Use It?

- A massive, online community with 26.4 million monthly active users
- 853,824 subreddits
- 150,000 communities
- 217 Countries
- 18 Billion Monthly Pageviews
- On track to take in $119 million in revenue for 2019

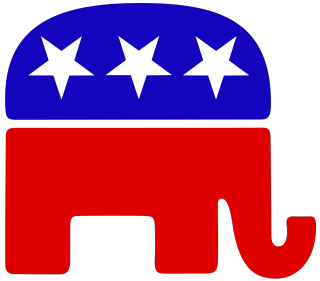
Acquiring and Processing the Data

1. Pull the posts using Reddit API

2. Amending the Stop Word List

3. Binarizing the Target: Democrats : 1 and Conservative : 0

Breaking down the Top Words



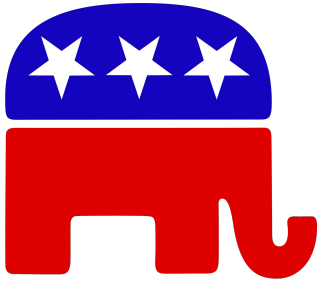
/R/Conservative

the	437
to	338
of	261
and	192
in	185
for	153
is	141
it	137

/R/Democrats

the	1028
to	752
trump	521
of	476
and	465
in	376
is	296
for	285

Breaking down the Top Words



/R/Conservative

biden	52
new	46
white	30
mueller	27
state	25
abortion	25
women	24
black	24

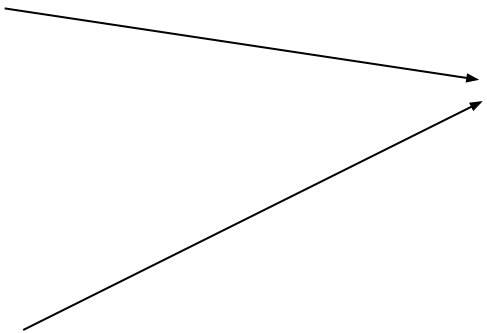
/R/Democrats

president	105
mueller	85
2020	83
house	79
new	78
report	73
white	70
twitter	64

Modeling the Data

- ***Models Tested:***
  - Logistic Regression
  - Multinomial Naive-Bayes
  - Random Forest

- ***Vectorizers Used***
  - CountVectorizer
  - TF-IDF



Optimal Model Setup:

CountVectorizer with Random Forest Classifier

Conclusions & Recommendations

- Potential for use in other online environments

- Additional steps: Sentiment Analysis, Refine Stop Words

- Inherent difficulty in defining similar classes: Conservatives and Democrats