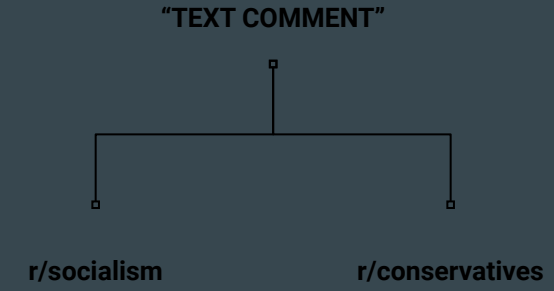
Natural Language Processing on Reddit

By Clay Carson

The Scope

1. Evaluate **sentiment** differences between the language of two subreddits:
    a. r/conservatives
    b. r/socialism

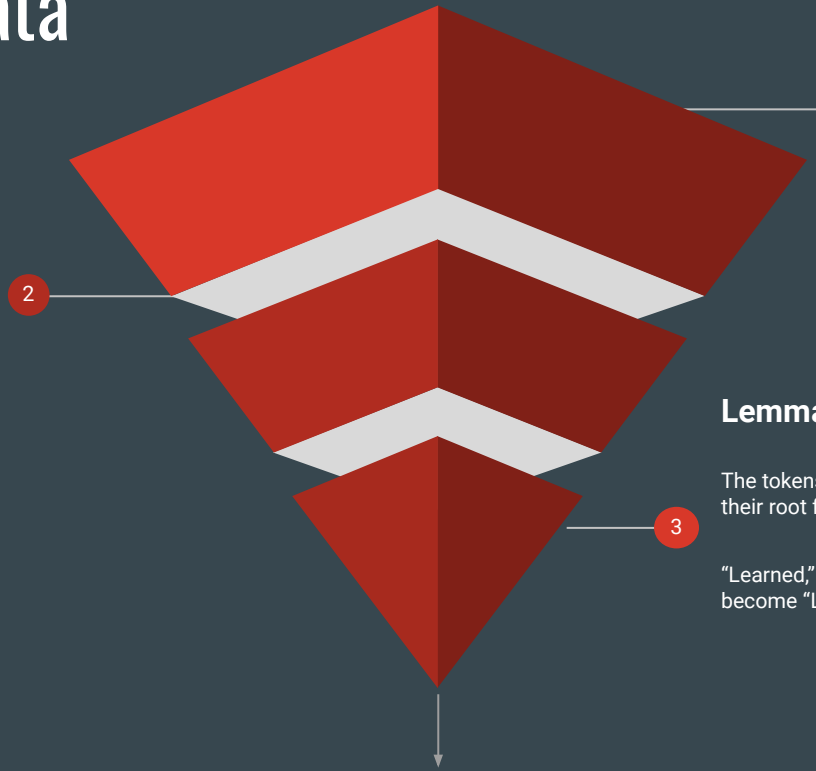
2. Create a model that can accurately **predict** to which subreddit a particular comment belongs



Cleaning the data

Scrape Comments

The first step is to acquire the comments from each subreddit via web scraping. The comments are then put into a dataframe and classified by the subreddit to which they belong.



“Tokenze” Comments

The comments are split into their individual words, or “tokenized.”

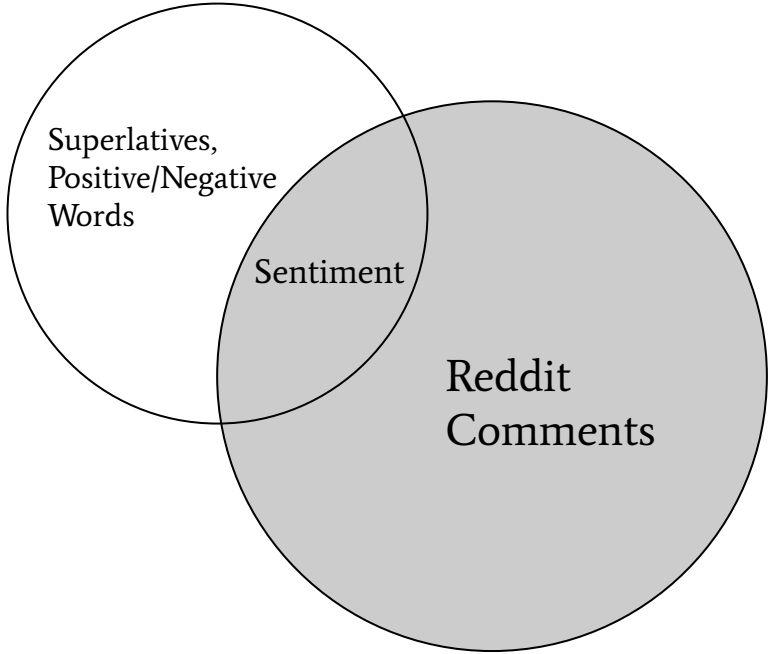
Lemmatize Tokens

The tokens are then shortened to their root forms using lemmatization.

“Learned,” “learning,” and “learns” all become “Learn.”

Sentiment Analysis

1. Three lists of words are scraped from the internet:
    a. Superlatives
    b. Positive Words
    c. Negative Words
2. The lemmatized words are compared to all three lists to determine overlap
3. The overlapping words in the reddit comments are counted and compared to the comments as a whole



Sentiment Analysis Results

	r/socialism	r/conservatives
Superlative Rate	0.117%	0.121%
Positive Word Rate	3.40%	3.27%
Negative Word Rate	1.88%	2.19%

Cloud Computing and Modeling

Lemmatized Words
+
Classification

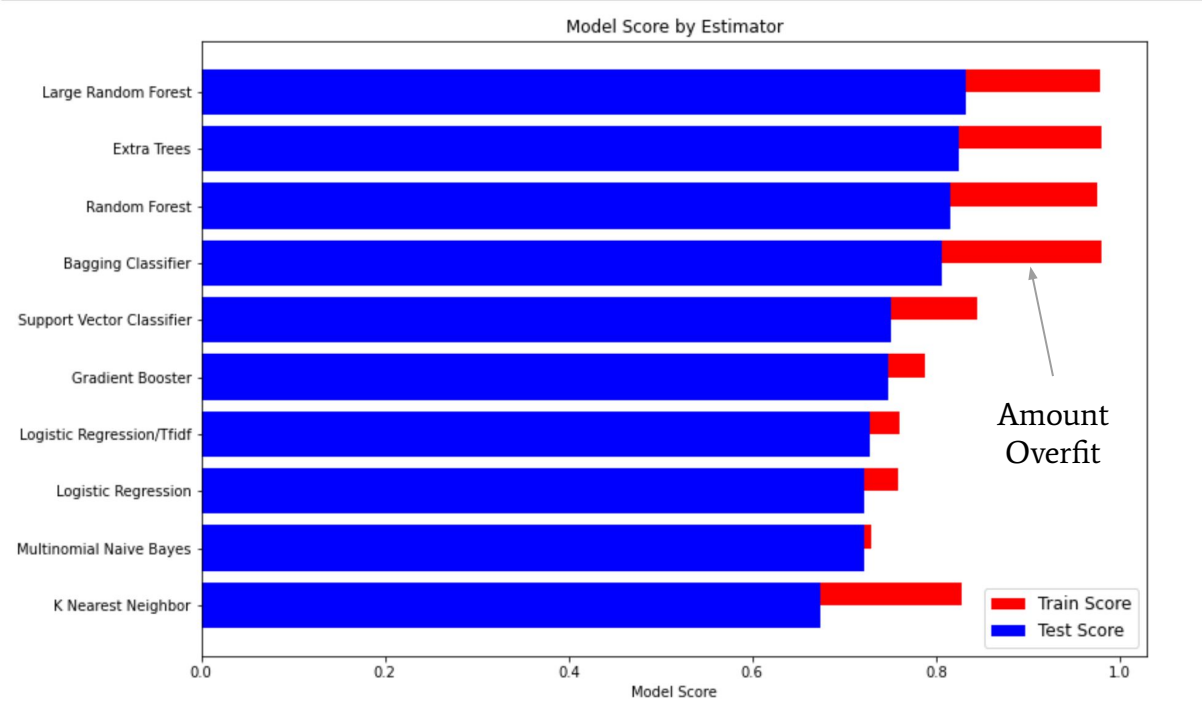
Predictive Model

Model Comparison

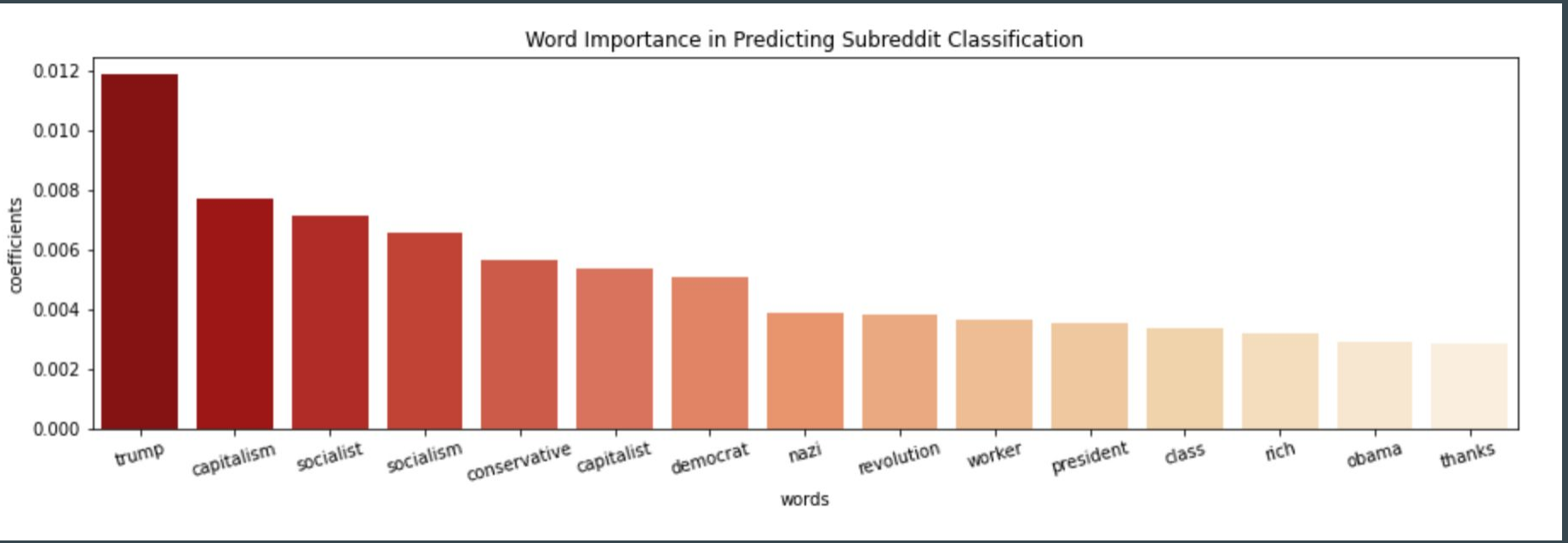
1,182 Models analyzing 40,000 comments in 45 minutes and 30 seconds

Best Model: Random Forest

Predictive Accuracy: **83.2%**



Most Important Words in Classification



Thanks for listening.