Reddit, White, and Blue

Can Natural Language Processing Identify Political Affinity?

DSIR-1019 Project 3
Prepared and Presented by
Peter Yonka

Why we're here...

A research firm with a political polling division needs to create a process to identify candidates for an online tracking study it's launching in six months. The study will require a steady stream of respondents with certain political affinities (Democrat, Republican, and Libertarian) to meet the sample requirements of the study.

Goal #1: Find a way to identify potential candidates with specific political affinities via several online resources (social media, survey responses, etc.).

Goal #2: Reveal the key words or phrases that aid in identifying people's political affinity.

With so much text-based data, can machine learning utilizing natural language processing help us achieve these goals?

Why Reddit?

Selected Subreddits:

- r/democrats: 158k Members
- r/Republican: 163k Members
- r/Libertarian: 442k Members

Data source: '109 Ridiculous Reddit Statistics & Facts to Know in 2020' (*https://websitebuilder.org/blog/reddit-statistics/)*

Getting ready...

A few of the cleaning steps:

- Pulled 3,000 submissions per subreddit (no duplicates)
- Cleaned out url data
- Combined title and selftext data
- Created text length and word count
- Prescreened word vectors, bigrams, and trigrams

Subreddit	# of Submissions (after cleaning)	% of Total Submissions
Republicans	2,987	33.8%
Democrats	2,984	33.8%
Libertarians	2,857	32.4%

Many models entered, one model emerged...

Model	Model Accuracy Score	Training Data Accuracy Score	Testing Data Accuracy Score
Logistic Regression	54.0%	64.7%	54.6%
Support Vector Machines	53.6%	81.7%	54.6%
Random Forest	49.3%	51.9%	46.9%
Gaussian Naive Bayes	48.0%	56.6%	49.9%

Many models entered, one model emerged...

Model	Model Accuracy Score	Training Data Accuracy Score	Testing Data Accuracy Score
Logistic Regression	**54.0%**	**64.7%**	**54.6%**
Support Vector Machines	53.6%	81.7%	54.6%
Random Forest	49.3%	51.9%	46.9%
Gaussian Naive Bayes	48.0%	56.6%	49.9%

Many models entered, one model emerged...

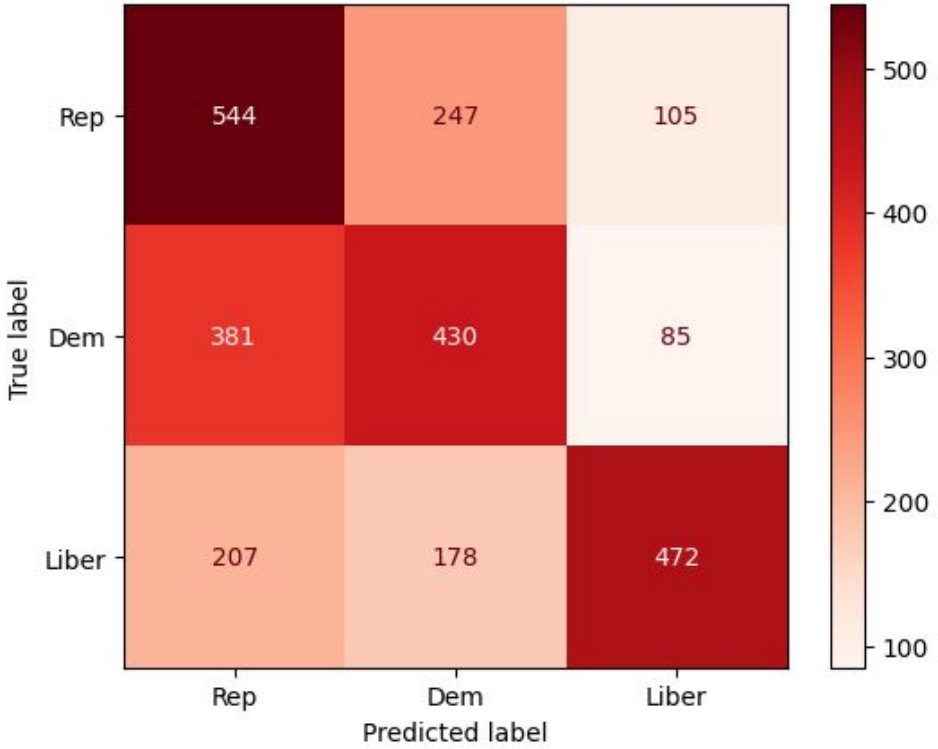
54.6% - 33.8% = 20.8% improvement

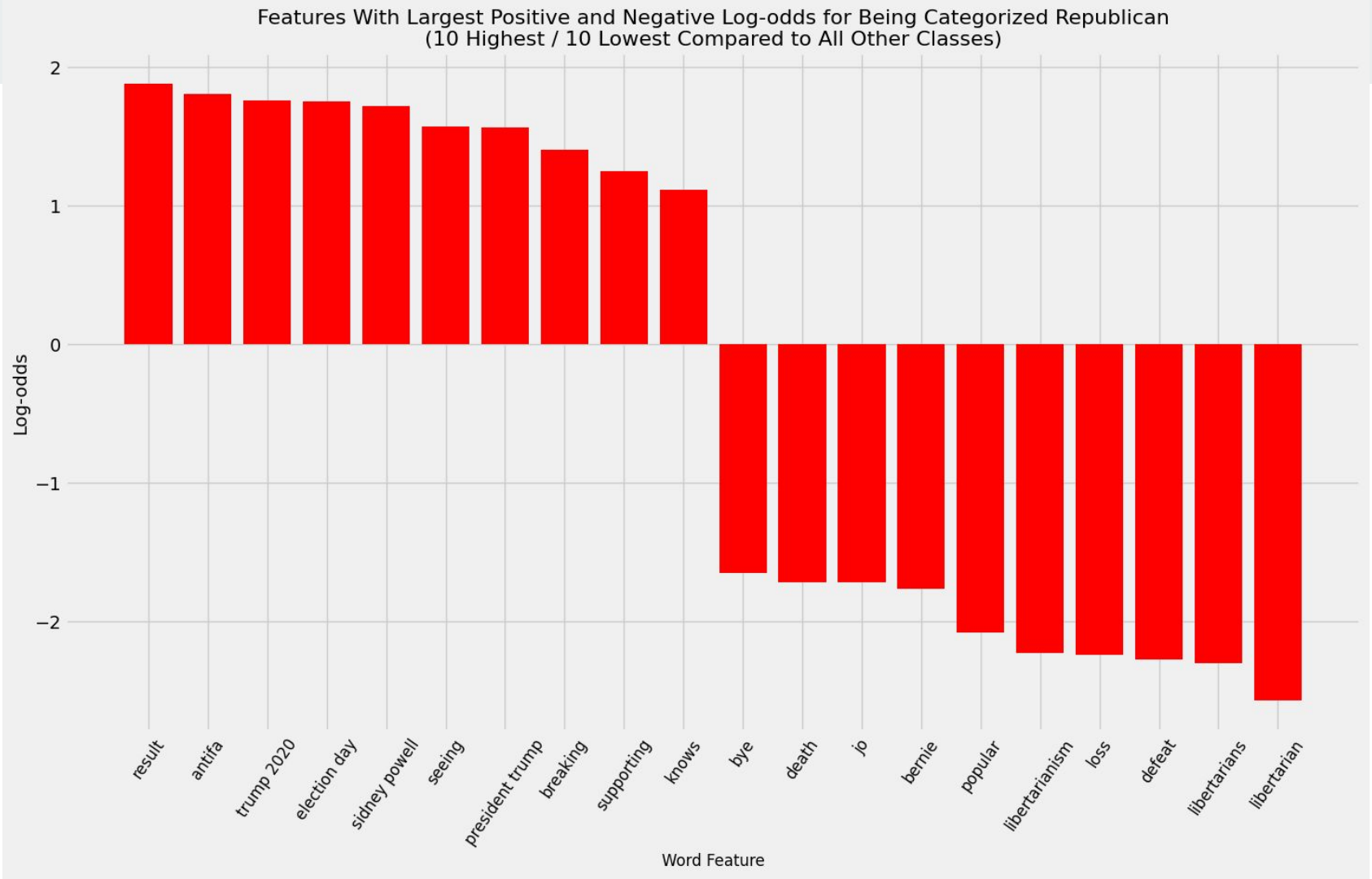
Model	Model Accuracy Score	Training Data Accuracy Score	Testing Data Accuracy Score
Logistic Regression	**54.0%**	**64.7%**	**54.6%**
Support Vector Machines	53.6%	81.7%	54.6%
Random Forest	49.3%	51.9%	46.9%
Gaussian Naive Bayes	48.0%	56.6%	49.9%

Where did we go right (and where did we go wrong)?

Confusion Matrix

- Run only on test data (data the model hadn't seen)
- Shows where misclassification are occuring

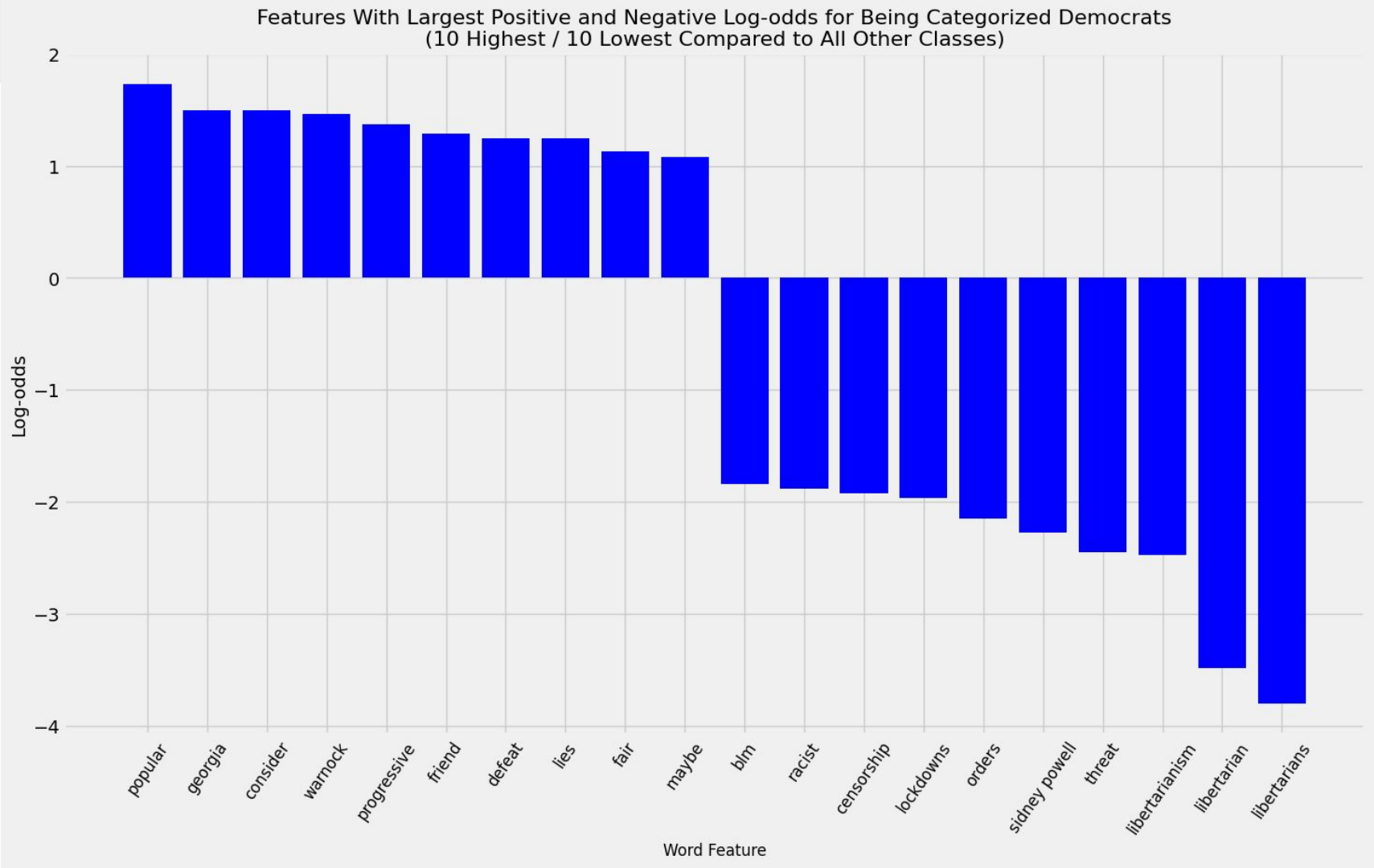



Features With Largest Positive and Negative Log-odds for Being Categorized Republican
(10 Highest / 10 Lowest Compared to All Other Classes)
Log-odds
2
1
0
−1
−2
result
antifa
trump 2020
election day
sidney powell
seeing
president trump
breaking
supporting
knows
bye
death
jo
bernie
popular
libertarianism
loss
defeat
libertarians
libertarian
Word Feature

Republican – features with highest and lowest odds

Subreddit	Largest Beta (feature = 'result')	Smallest Beta (feature = 'libertarian')
Republicans	**6.53**	**0.08**

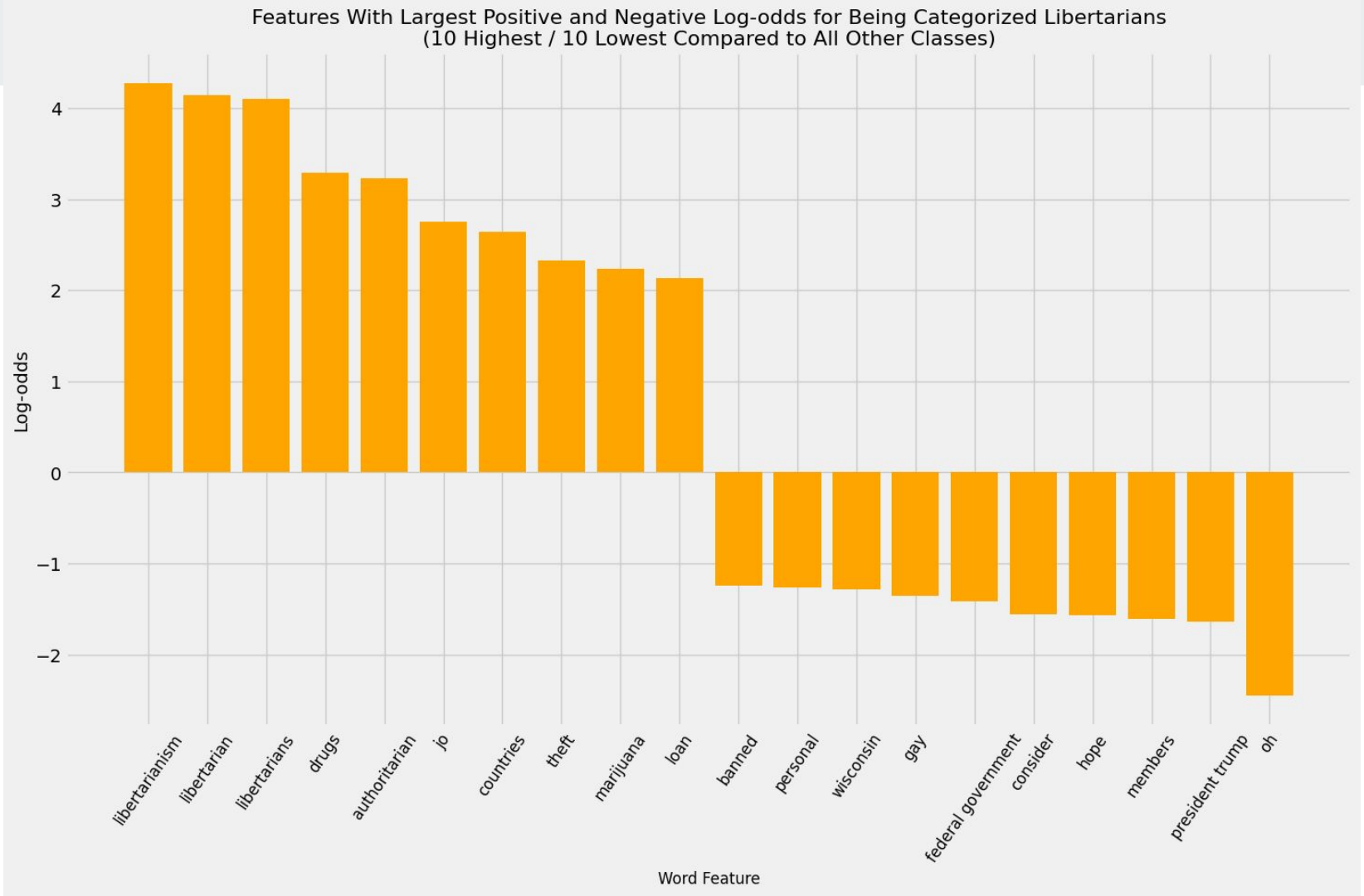
- For every single occurrence of the word 'result', the submission is 6.53x as likely to be in the Republican subreddit compared to all other classes, holding all else constant.
- For every single occurrence of the word 'libertarian', the submission is 0.08x as likely to be in the Republican subreddit compared to all other classes, holding all else constant.


Features With Largest Positive and Negative Log-odds for Being Categorized Democrats
(10 Highest / 10 Lowest Compared to All Other Classes)
Log-odds
2
1
0
−1
−2
−3
−4
popular
georgia
consider
warnock
progressive
friend
defeat
lies
fair
maybe
blm
racist
censorship
lockdowns
orders
sidney powell
threat
libertarianism
libertarian
libertarians
Word Feature

Democrats – features with highest and lowest odds

Subreddit	Largest Beta (feature = ‘popular’)	Smallest Beta (feature = ‘libertarians’)
Democrats	**5.65**	**0.02**

- For every single occurrence of the word ‘popular’, the submission is 5.65x as likely to be in the Democrats subreddit compared to all other classes, holding all else constant.
- For every single occurrence of the word ‘libertarians’, the submission is 0.02x as likely to be in the Democrats subreddit compared to all other classes, holding all else constant.


Features With Largest Positive and Negative Log-odds for Being Categorized Libertarians
(10 Highest / 10 Lowest Compared to All Other Classes)
Log-odds
4
3
2
1
0
−1
−2
libertarianism
libertarian
libertarians
drugs
authoritarian
jo
countries
theft
marijuana
loan
banned
personal
wisconsin
gay
federal government
consider
hope
members
president trump
oh
Word Feature

Libertarians – features with highest and lowest odds

Subreddit	Largest Beta (feature = ‘libertarianism’)	Smallest Beta (feature = ‘oh’)
Libertarians	**71.66**	**0.09**

- For every single occurrence of the word ‘libertarianism’, the submission is 71.66x as likely to be in the Libertarians subreddit compared to all other classes, holding all else constant.
- For every single occurrence of the word ‘oh’, the submission is 0.09x as likely to be in the Libertarians subreddit compared to all other classes, holding all else constant.

Next Steps

- Assess model effectiveness over time
  - Too general: perhaps lower predictive power
  - Too specific: would need to be updated with every news cycle
- Additional model tuning
  - Improving bias/variance tradeoff
  - Analysis of missed predictions
- Additional preprocessing work
  - Lemmatizing and stemming
  - Custom stop words
- Utilizing numeric data with the vectorized word data
  - Word counts
  - Text length
  - Scores
  - Sentiment analysis

Thank you!

Please let me know if you have any questions.