Campaign Selection by Reddit Comment Analysis

Griffin Talan

Data Science Problem

A local political donor wants to know which liberal politicians are trending among Reddit users so that they can be most effective in their donations

- Who is the most popular candidate?
- What trends exist among the most popular candidates?

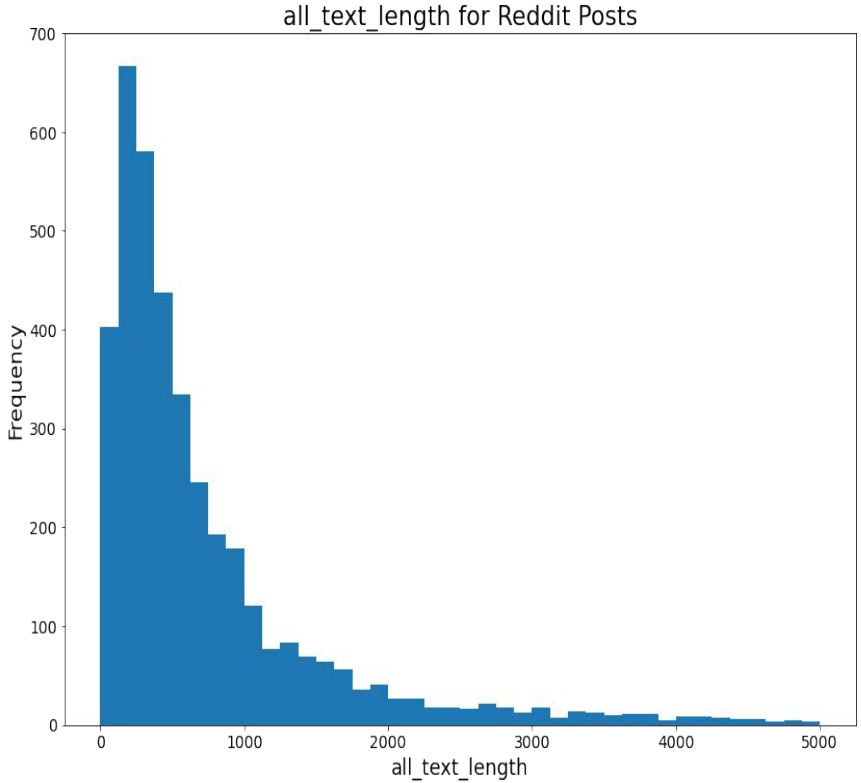
Agenda

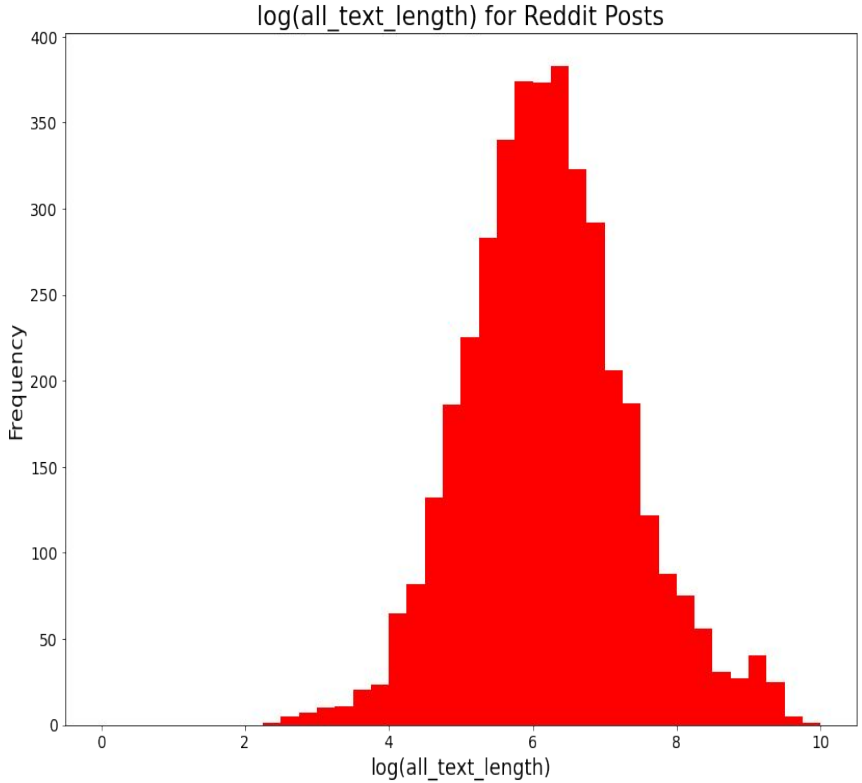
- Data Processing and EDA
- Model Selection and Evaluation
- Interpretation of Model Coefficients
- Donation Suggestions
- Next Steps

Data Processing and EDA

- Removed [deleted] and [removed] and blank seltext records
- Converted created_utc column into days_old column
- Combined title and seltext into alltext
- Created text length fields
- Took the logarithm of the all_text_length field
- Pulled 2000 ‘good’ records from each subreddit

log(all_text_length)



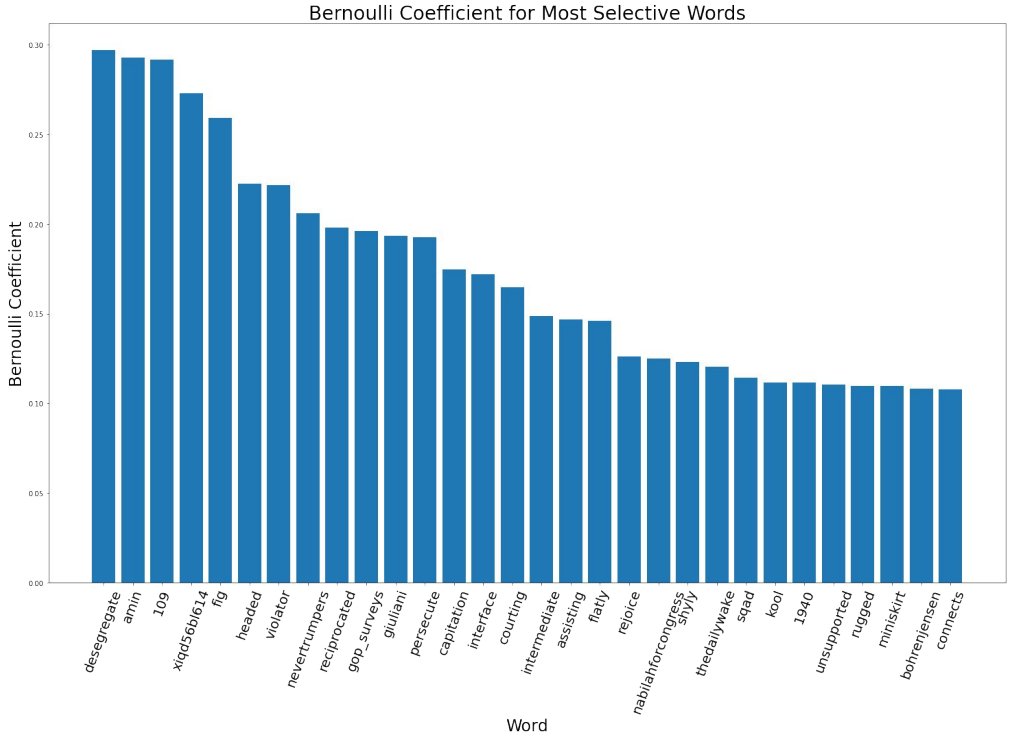


Model Selection - Naive Bayes Bernoulli Classifier

```
--------------------All Text--------------------
Baseline Accuracy: 0.5
Train Accuracy: 0.7897
Test Accuracy: 0.585

True Positives: 393
True Negatives: 192
False Positives: 308
False Negatives: 107

Specificity: 0.384
Sensitivity: 0.786

{'nb__alpha': 0.01, 'tfidf__lowercase': False, 'tfidf__ngram
_range': (1, 1), 'tfidf__stop_words': 'english'}
```

- TfidfVectorizer
- Lack of independence between features caused limitations
- Much better at predicting conservative subreddit posts

Candidate Choice - Naive Bayes Bernoulli Classifier



Important Names

- Non-Democrats
  - Rudy Giuliani (giuliani)
  - Michel Platini (platini)

- Democrats
  - Nabilah Islam (nabilahforcongress)
  - Jensen Bohren (bohrenjensen)
  - Bob Moser (moser)
  - Claire McCaskill (mccaskill)
  - Darren Soto (darren)
  - Tomas Ramos (tomasforcongress)

Non-Democrats

- Rudy Giuliani (giuliani)
  - Only republican

- Michel Platini (platini)
  - Former VP of FIFA
  - Caught up in ‘Qatargate’

Democrats

- Nabilah Islam (nabilahforcongress)
  - 2020 House Cand, Georgia
- Jensen Bohren (bohrenjensen)
  - Lost 2020 Senate Mississippi
- Bob Moser (moser)
  - Wrote “Blue Dixie: Awakening the South’s Democratic Majority”
- Claire McCaskill (mccaskill)
  - Former Missouri Senator
- Darren Soto (darren)
  - 2020 House Cand, Florida
- Tomas Ramos (tomasforcongress)
  - 2020 House Cand, New York

Campaign Donation Suggestions

	bernoulli_coefficient
giuliani	0.193360
nabilahforcongress	0.124922
bohrenjensen	0.108311
platini	0.087049
moser	0.061800
mccaskill	0.061135
darren	0.048511
tomasforcongress	0.045189

- Invest in Nabilah Islam
- Invest in the southern states

Next Steps - Support Vector Machine Classification

```
------All Text------
Baseline Accuracy: 0.5
Train Accuracy: 0.8233
Test Accuracy: 0.785

True Positives: 456
True Negatives: 329
False Positives: 155
False Negatives: 60

Specificity: 0.6798
Sensitivity: 0.8837
```

- Use a high-performing ‘black box’ model to find subreddits that are politically aligned for campaign marketing research purposes

- Custom Support Vector Machine with TfidfVectorizer/StandardScaler
  - Parameters from best GridSearch SVM
  - Separately scaled text and numeric data