

Republican or Democrat

Which Subreddit is it?

reddit.com/r/Republican/new/

LOG IN SIGN UP

Search

LOG IN

Search

JOIN

JOIN

Hot New Top ...

Posted by u/Yosoff First Principles 5 minutes ago

Here you go @JoeBiden!

twitter.com/realDo... ↗

Donald J. Trump ✅
@realDonaldTrump

Here you go @JoeBiden!

About Community

/r/Republican is a partisan place for Republicans to interact with other Republicans.

154k Members 942 Online

Created Oct 10, 2008

D Democrats: Unite for a Better Future

Posts Rules Register to Vote! Social Media

Hot New Top Today ...

1.1k

Posted by u/walter1950 13 hours ago 🎉

Watch Obama Absolutely Torch Trump During His Debut Campaign Event for Biden, no lies or Mistruths just the pure facts.

rollingstone.com/politi... ↗

37 Comments Share Save ...

PROMOTED · Posted by u/BoratSubsequentMovie 1 hour ago 🔒

After being cancelled in Kazakhstan, Borat is back to restore its glory. Stream his subsequent moviefilm now only on Prime Video.

About Community

The Democratic Party is fighting for a country where everyone, from every walk of life, has an equal chance at the American dream. This sub offers daily news updates, policy analysis, links, and opportunities to participate in the political process. We are here to get Democrats elected up and down the ballot.

147k Democratic voters 432 Online

Created Oct 4, 2008

Background

r/Republican

154k members

Republican group with most members

90,351 submissions from last 8 years

r/democrats

146k members

Democratic group with most members

123,668 submissions from last 8 years

Two Goals:

Predict the subreddit

Understand word appearances

Common?

Unique?

Frequencies?

Surprises?

Methodology

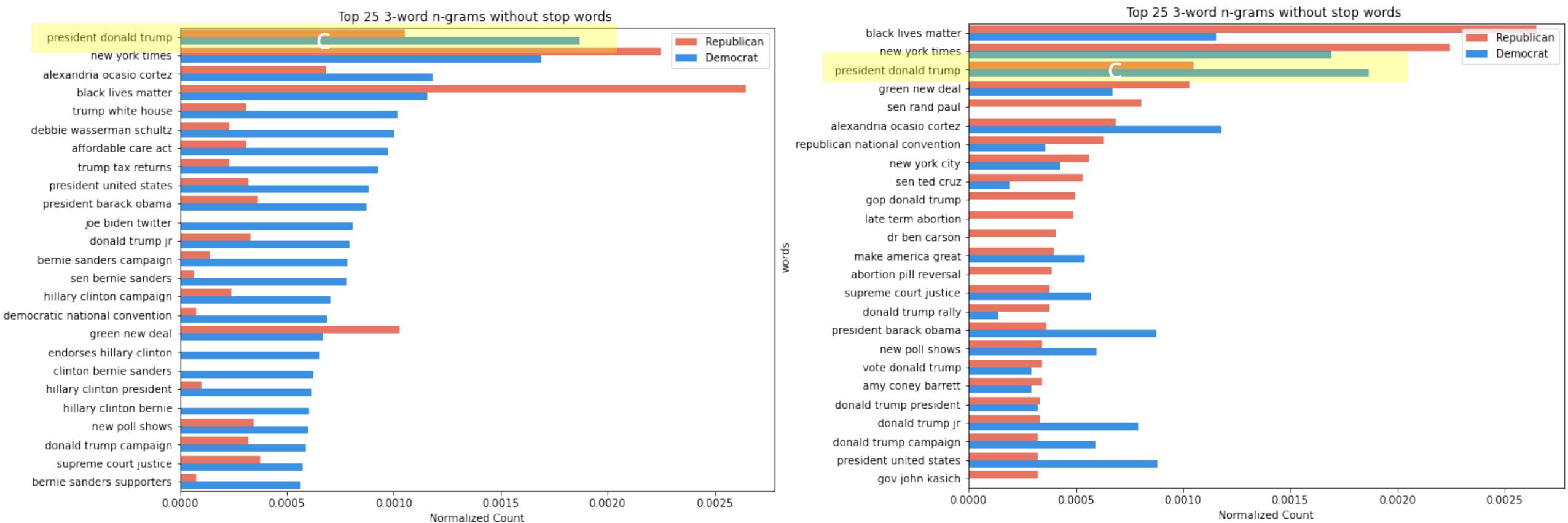
Understand
Word
Appearances

- ▶ Downloaded 8 yrs data from two subreddits
 - ▶ Over 90,000 submissions from each one
- ▶ Analyzed words in titles of submissions
 - ▶ Common, Unique, Tone

Predict the
Subreddit

- ▶ Analyzed urls
 - ▶ Common, Frequency
- ▶ Analyzed title word count, title length
- ▶ Created models to predict
- ▶ Reviewed prediction results
- ▶ Reviewed predictions to see if they matched previous analysis

Understand Word Appearances



Common Words

3 n-gram words

Took top 900 for each and compared

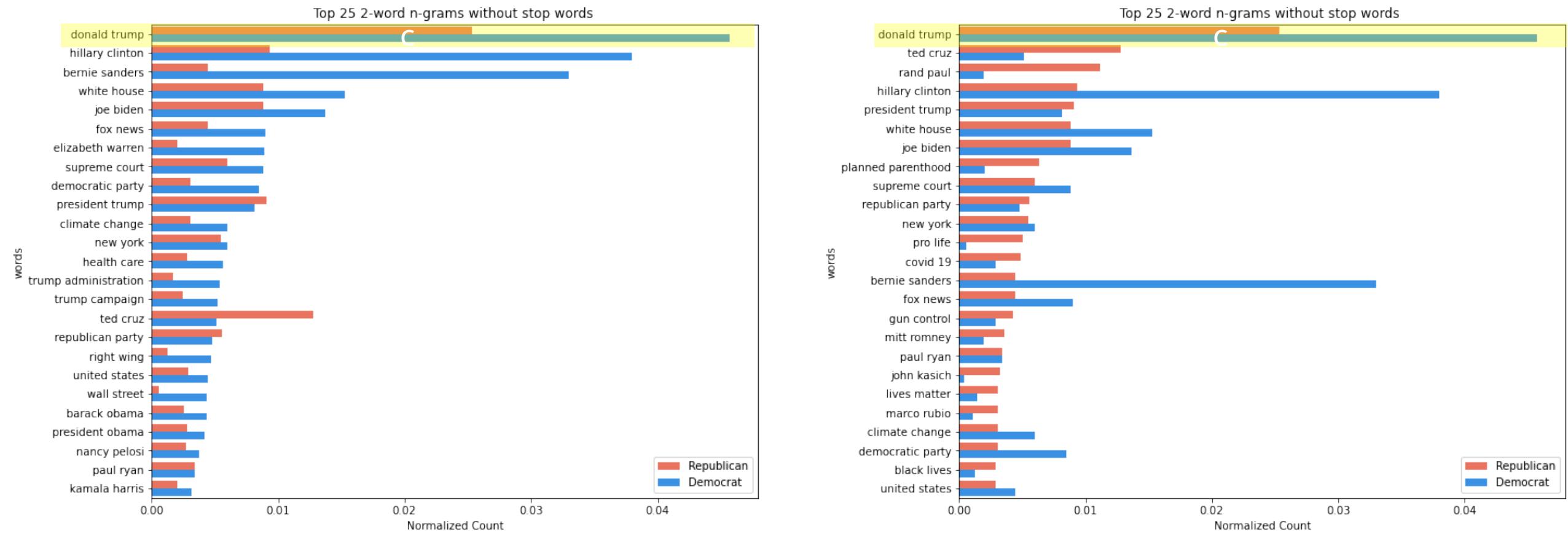
Does not include stop words

Normalized Counts

One of top words for both is president donald trump

Very dissimilar distributions - some are missing from one party

Mostly key names/causes in each party



Common Words

2 n-gram words

Took top 900 for each and compared

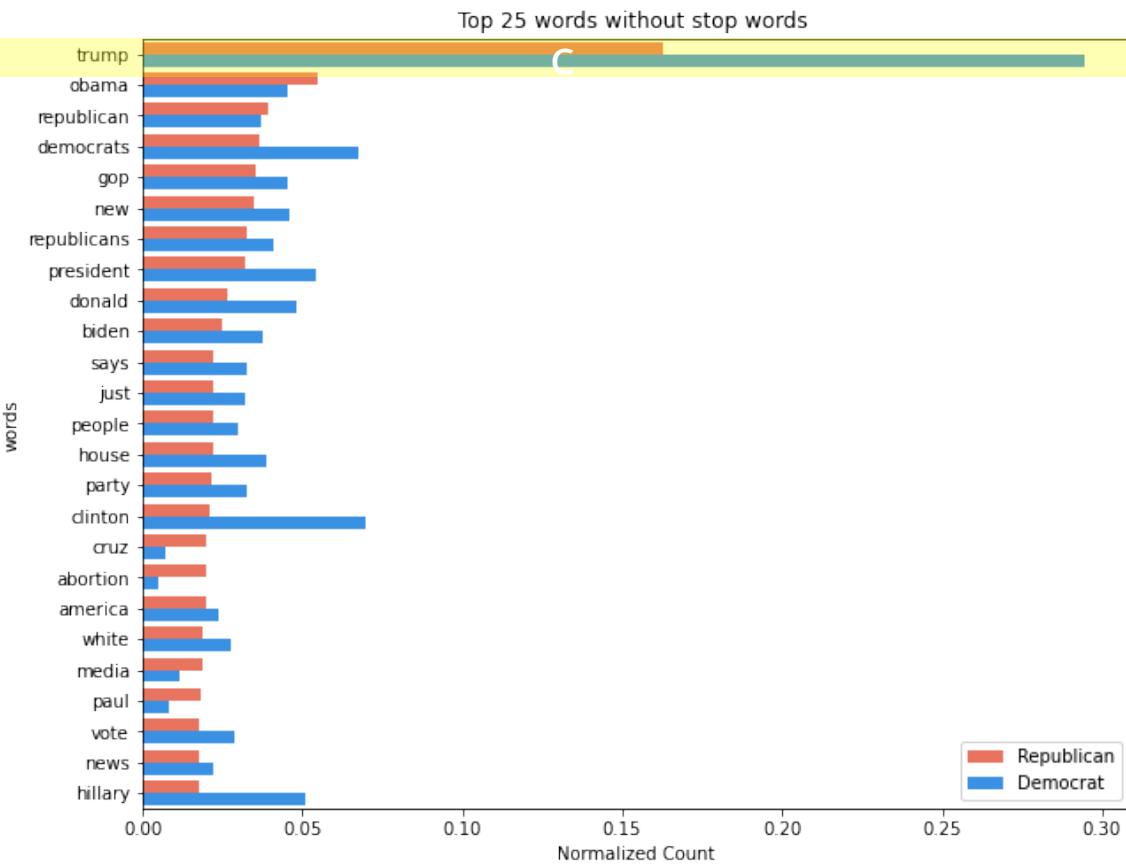
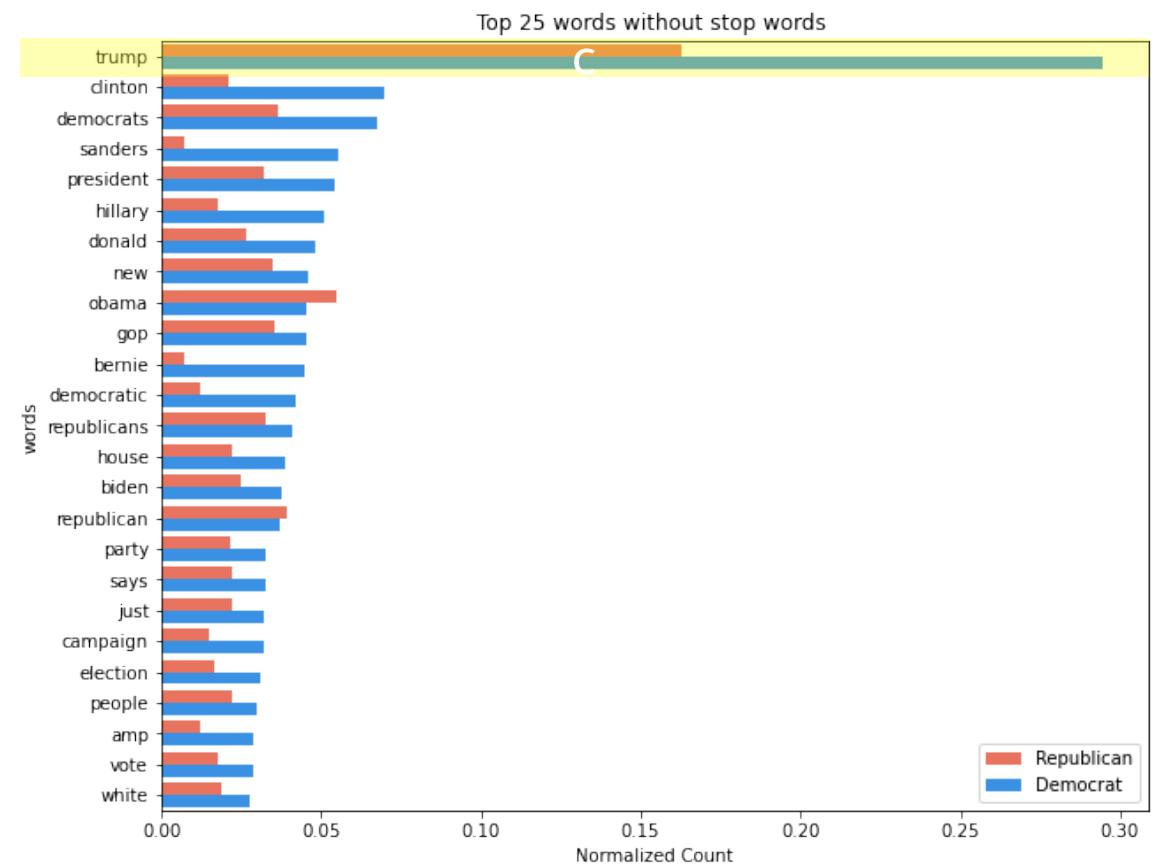
Does not include stop words

Normalized Counts

Top words for both is Donald trump, but this is disproportionately referenced by Democrat

Generally dissimilar distribution

Mostly key names in each party



Common Words

1 n-gram words

Took top 900 for each and compared

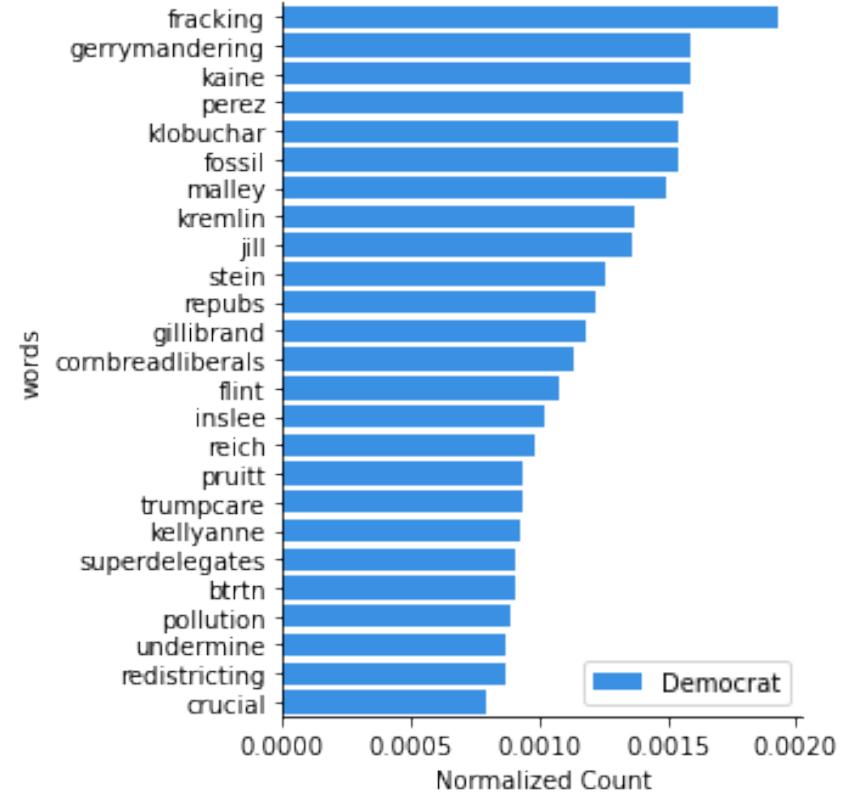
Does not include stop words

Normalized Counts

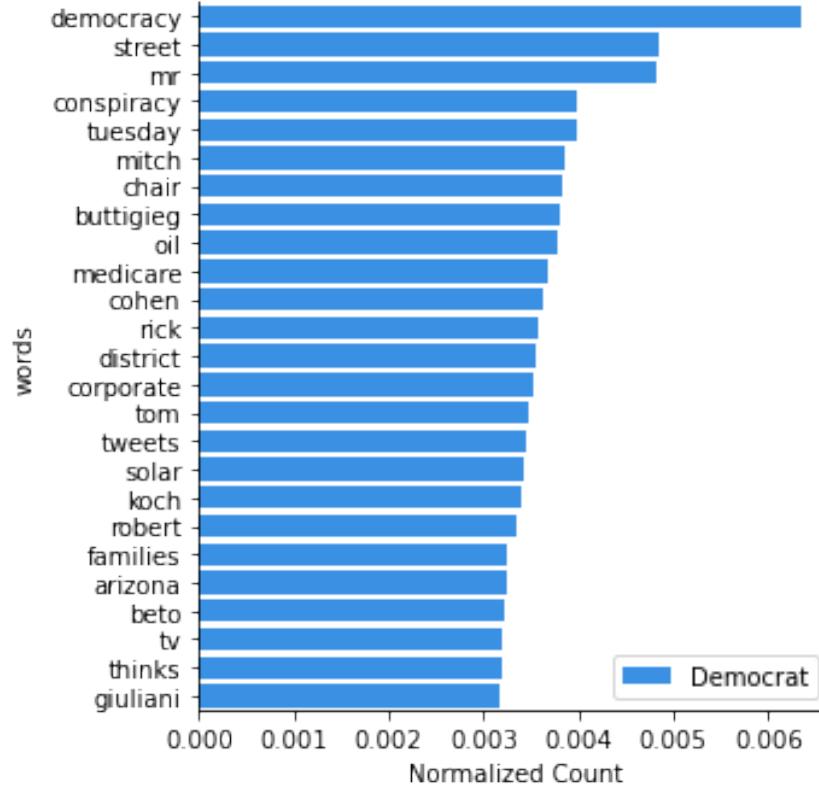
Top words for both is Trump, but this is disproportionately referenced by Democrat

Generally similar distribution

Top 25 unique words to a subreddit, of top 5000 words per each, ngram is (1, 1).



Top 25 unique words to a subreddit, of top 1000 words per each, ngram is (1, 1).



Unique Words

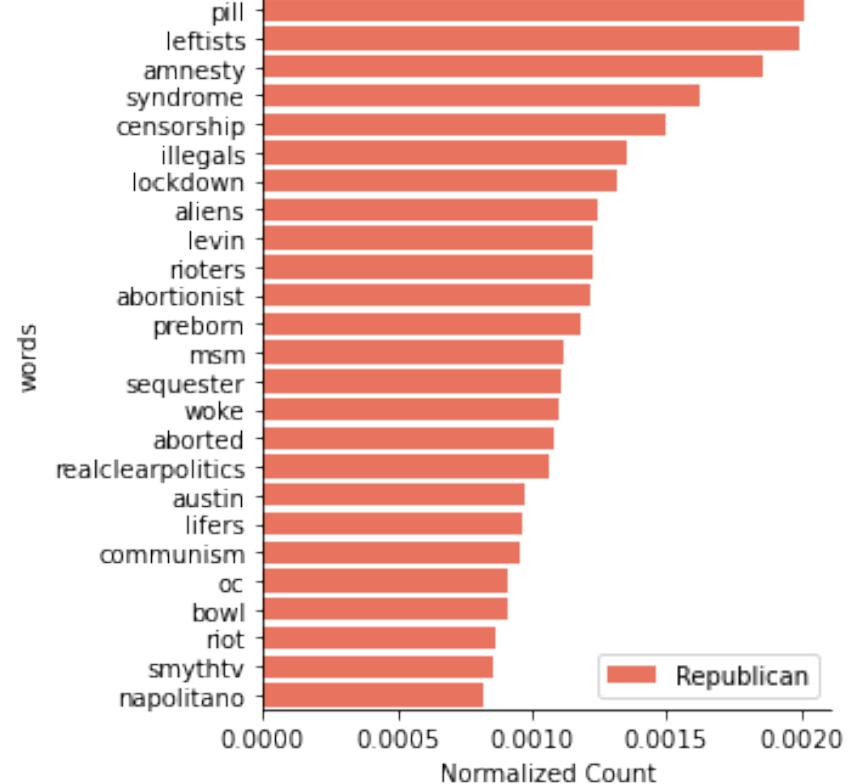
Took top 5000 and 1000 for each and compared

Includes stop words

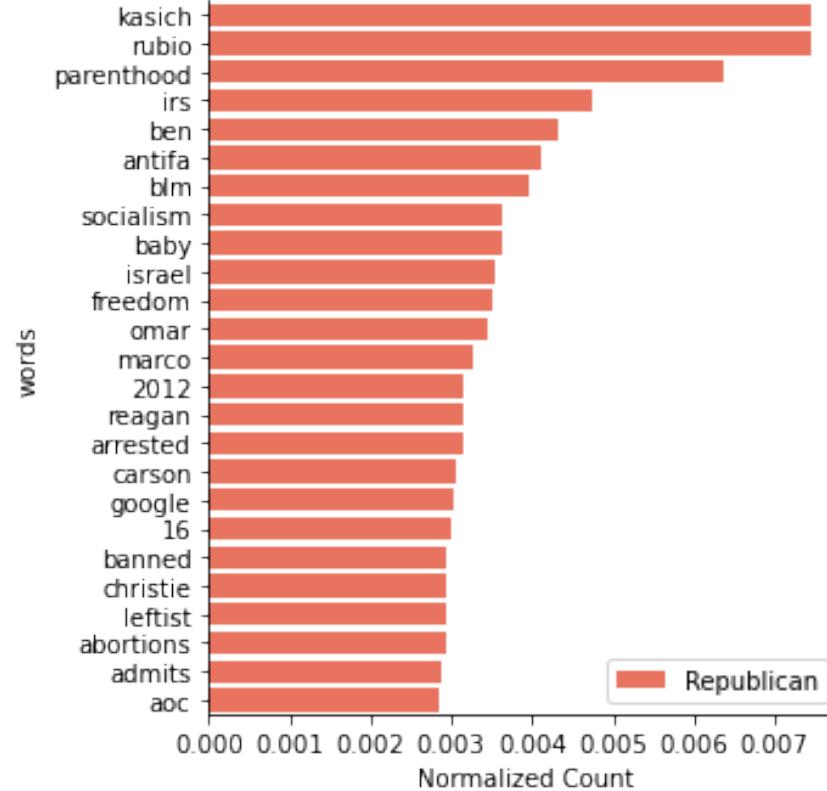
Normalized Counts

Terminology is interesting

Top 25 unique words to a subreddit, of top 5000 words per each, ngram is (1, 1).



Top 25 unique words to a subreddit, of top 1000 words per each, ngram is (1, 1).



Unique Words

Took top 5000 and 1000 for each and compared

Includes stop words

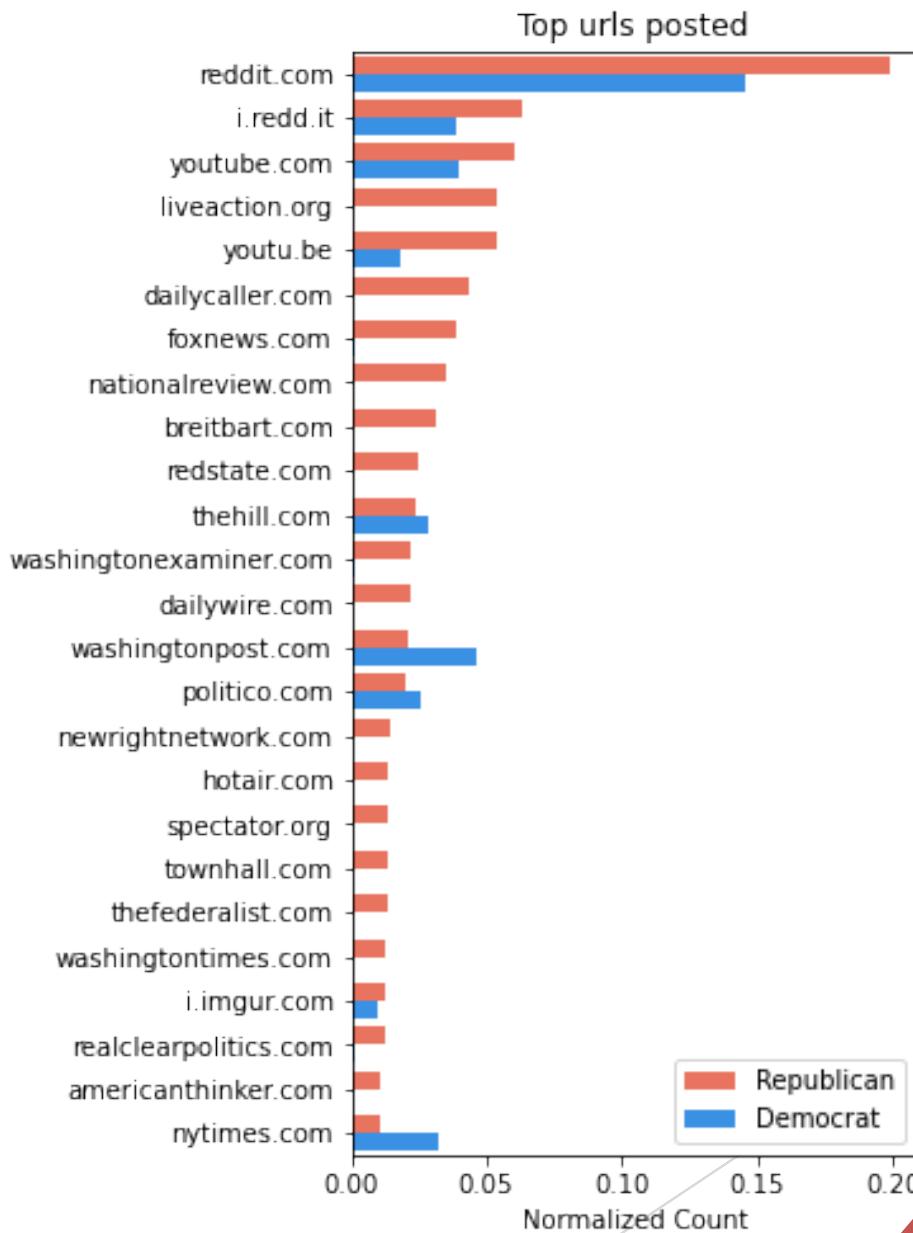
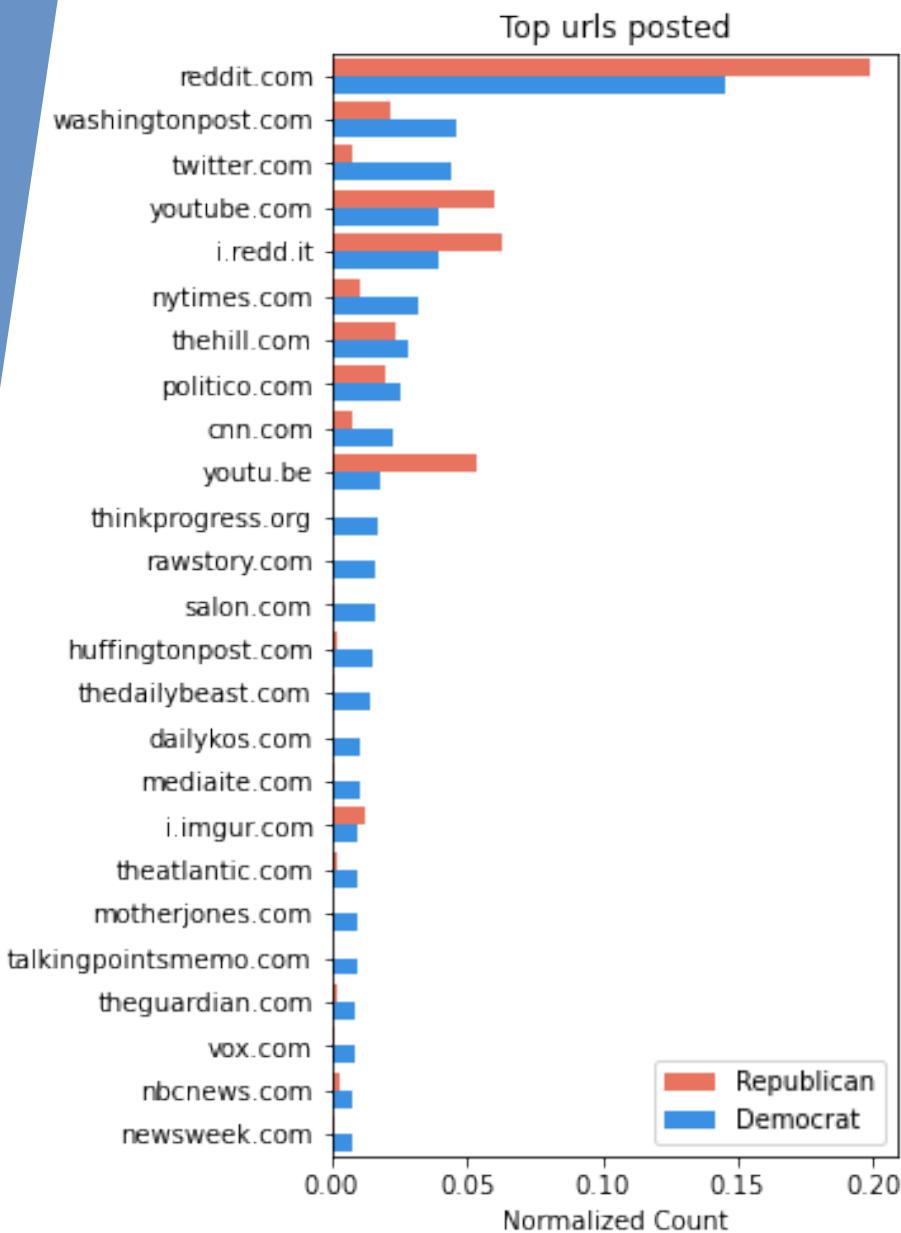
Normalized Counts

Terminology is interesting

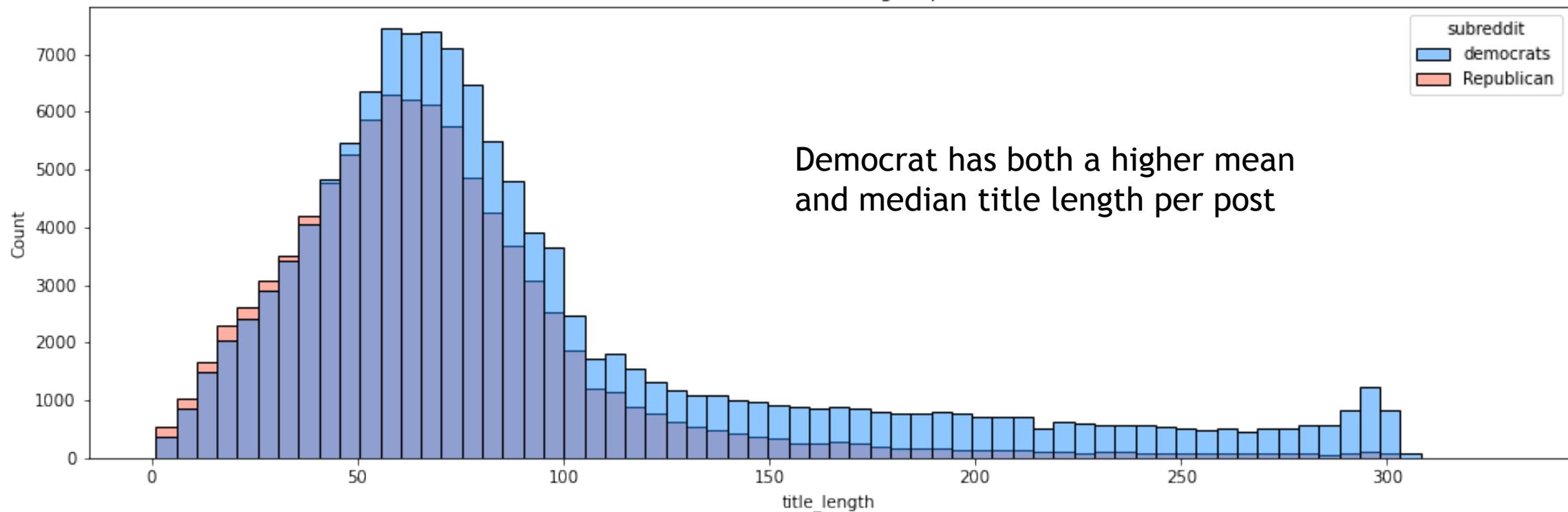
Understand Other Appearances

Mostly National News Organizations

URLS

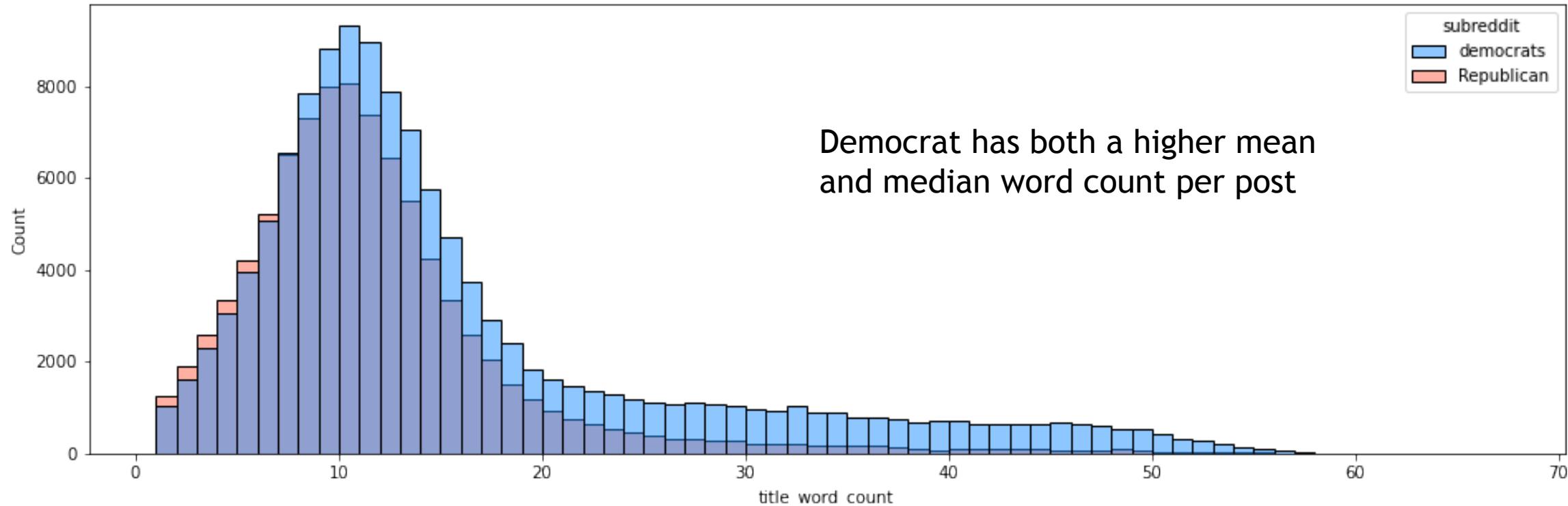


Distribution of Title Lengths per Post



	Democrat	Republican
Overall Count	123,668	90,351
Mean	95.6	70.0
Median	74	64

Distribution of Word Counts in Title per Post



	Democrat	Republican
Overall Count	123,668	90,351
Mean	15.6	11.3
Median	12	10

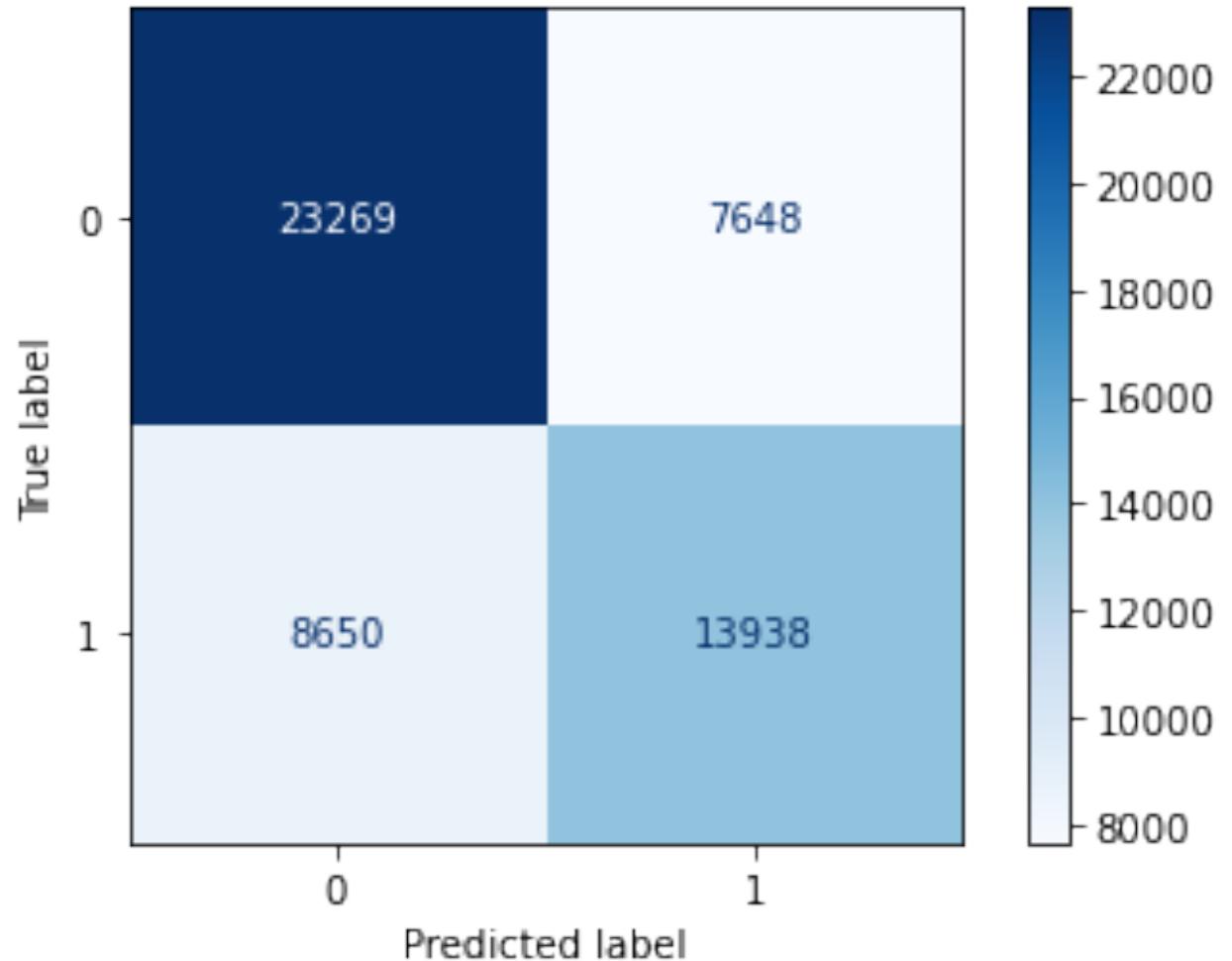
Predictions

Modeling Criteria:

- ▶ Performed a GridSearch for three types of Models
 - ▶ Multinomial Naïve Bayes
 - ▶ Logistic
 - ▶ RandomForestClassifier
- ▶ Tried two vectorizers:
 - ▶ CountVectorizer
 - ▶ Term Frequency-Inverse Document Frequency
- ▶ Also tried with and without urls, and with and without status length, title count

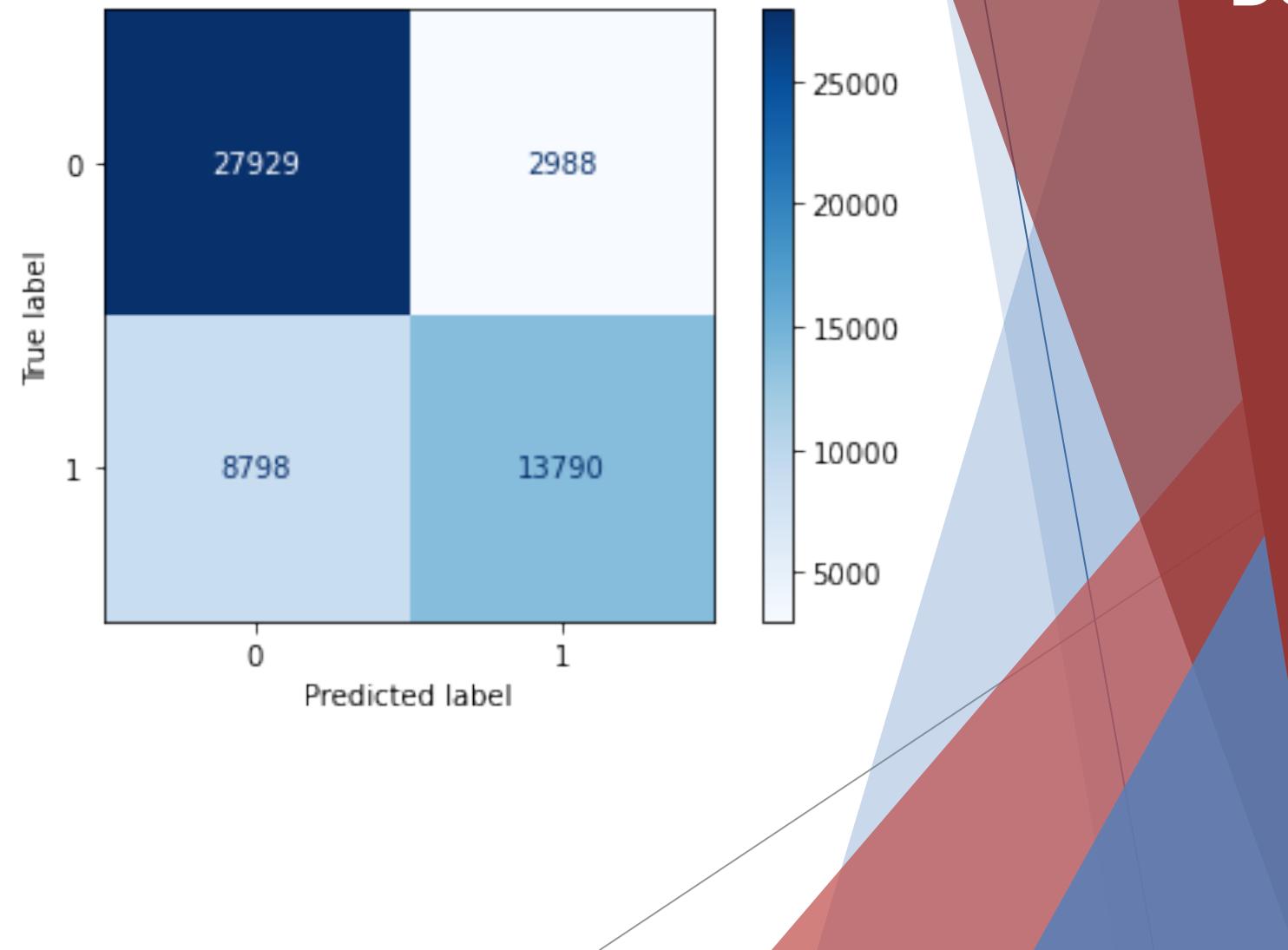
Best Model - NLP only:

- ▶ CVEC, Logistic Regression
- ▶ Baseline score:
 - ▶ 57.8% are Democrats
- ▶ Accuracy:
 - ▶ Cross-Val: 69.0%
 - ▶ Train: 71.4%
 - ▶ Test: 69.5%
- ▶ Sensitivity: 61.7%
- ▶ Specificity: 75.3%

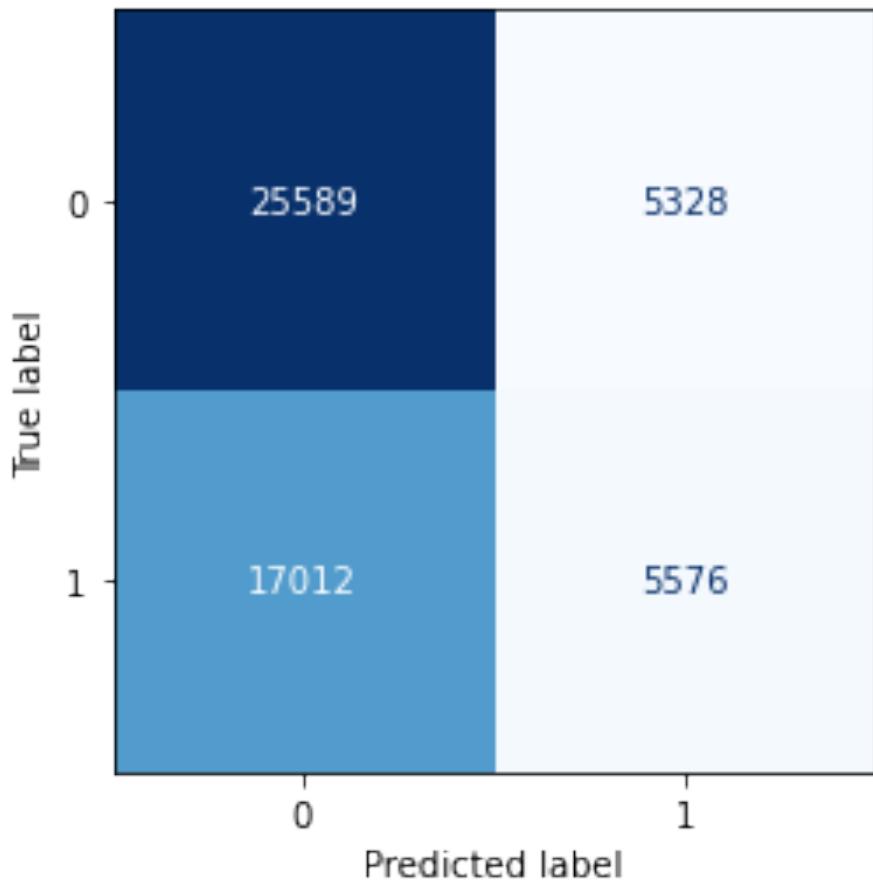


Best Model - URLs only:

- ▶ TVEC, Logistic Regression
- ▶ Baseline score:
 - ▶ 57.8% are Democrats
- ▶ Accuracy:
 - ▶ Cross-Val: 77.6%
 - ▶ Train: 79.7%
 - ▶ Test: 78.0%
- ▶ Sensitivity: 90.3%
- ▶ Specificity: 61.1%



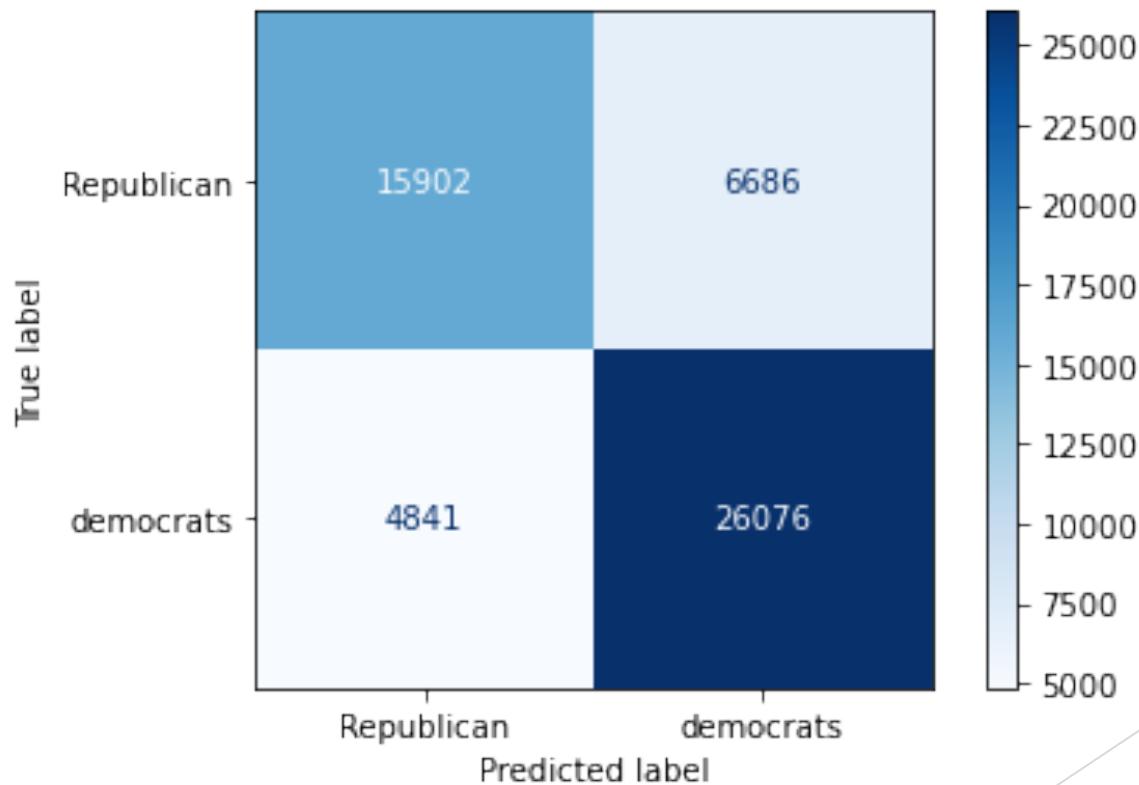
Model - Counts:

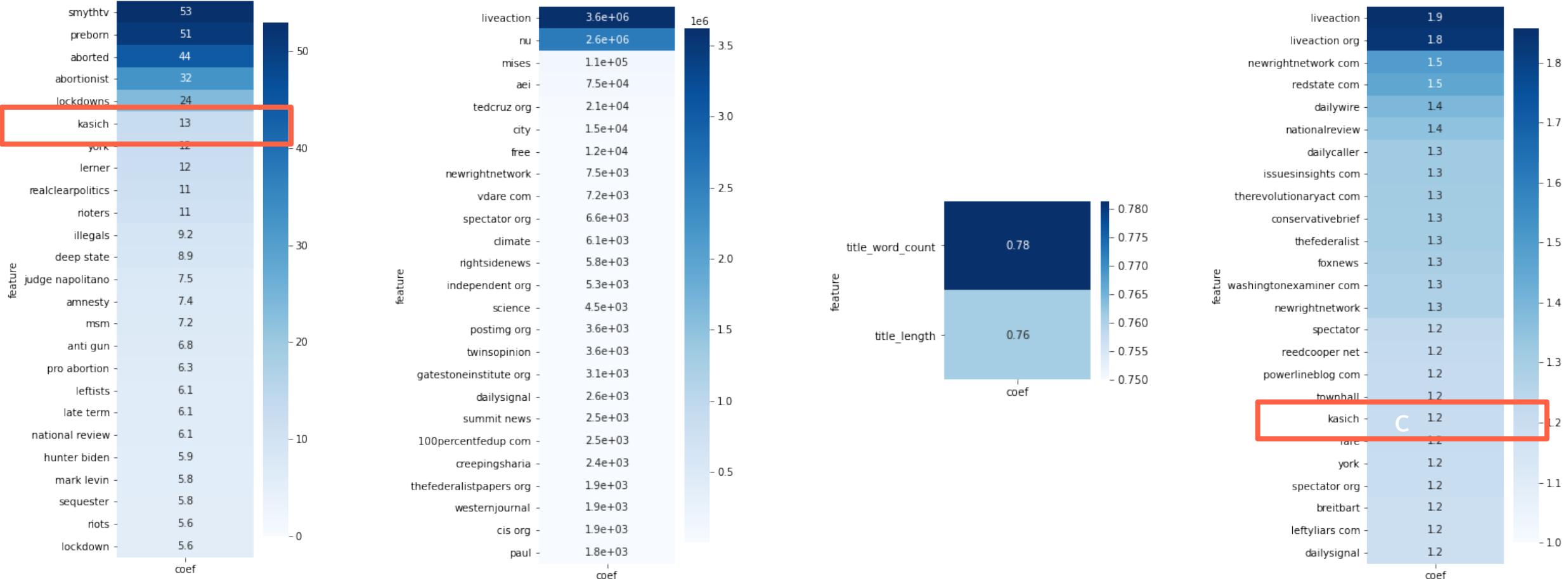


- ▶ Logistic Regression
- ▶ Baseline score:
 - ▶ 57.8% are Democrats
- ▶ Accuracy:
 - ▶ Cross-Val: XXX%
 - ▶ Train: 58.5%
 - ▶ Test: 58.2%
- ▶ Sensitivity: 82.8%
- ▶ Specificity: 24.7%

Best Model - Overall:

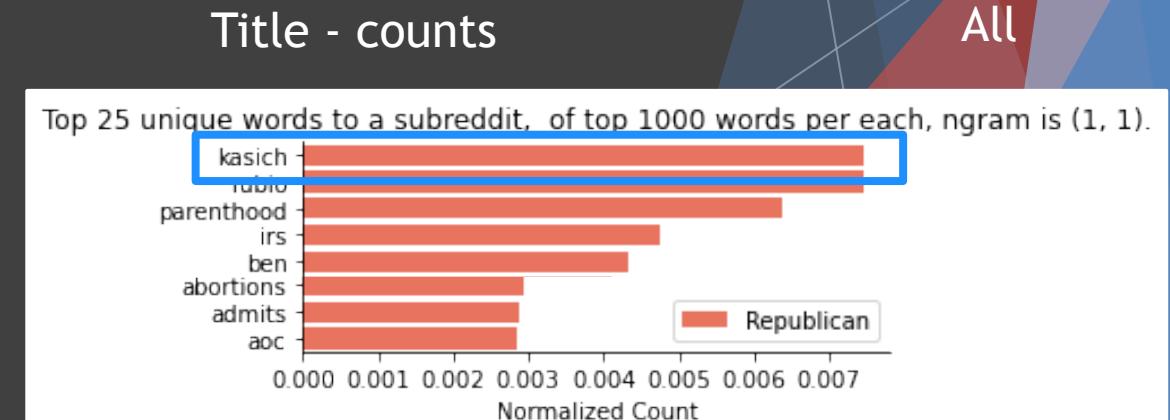
- ▶ Logistic Regression
- ▶ Baseline score:
 - ▶ 57.8% are Democrats
- ▶ Accuracy:
 - ▶ Train: 80.0%
 - ▶ Test: 78.5%
- ▶ Sensitivity: 70.4%
- ▶ Specificity: 84.3%





Review Features

Urls dominate the combined model



Conclusions

- ▶ ‘Trump’ is mentioned almost twice as often in Democrats vs Republicans
- ▶ Democrats like to write more words
- ▶Urls tell more about the subreddit than the words in the title
- ▶ Combining the words, Urls, and title description is the best

Next Steps

- ▶ Explore the relationships in the words more
- ▶ What ranking of words count for each least are the features in each model
- ▶ Explore similar subreddits - libertarian, conservative, progressive, liberal

VOTE!!!