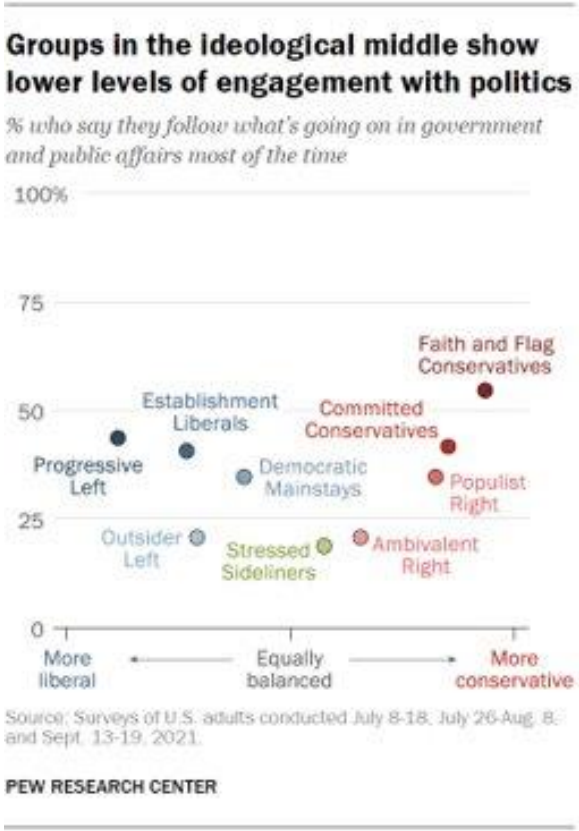
predicting reddit post origin with natural language processing

Introduction

The recent whistleblower incident from facebook has emphasized that dramatic content or ideologies of a vocal minority are disproportionally prevalent in the news, due to engagement.

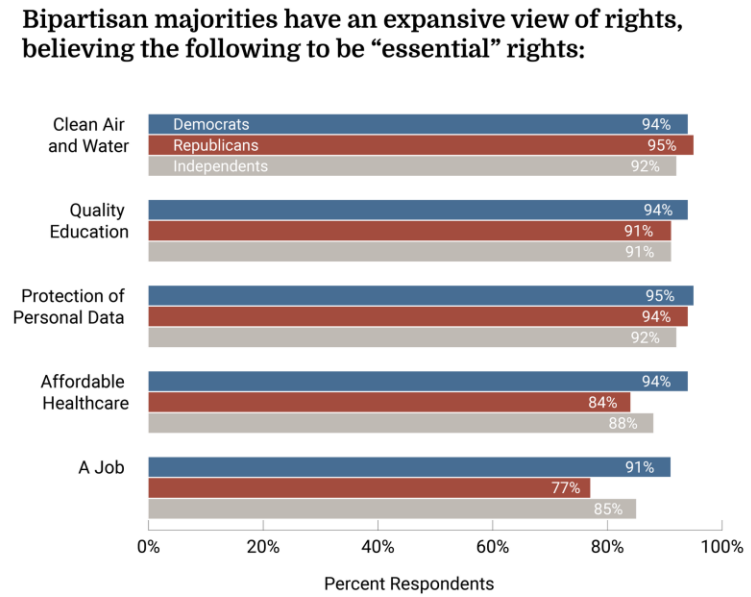
This contrast between media characterization and the average american may cause citizens to have misconceptions about others ideologies, and artificially contribute to political polarization



Introduction

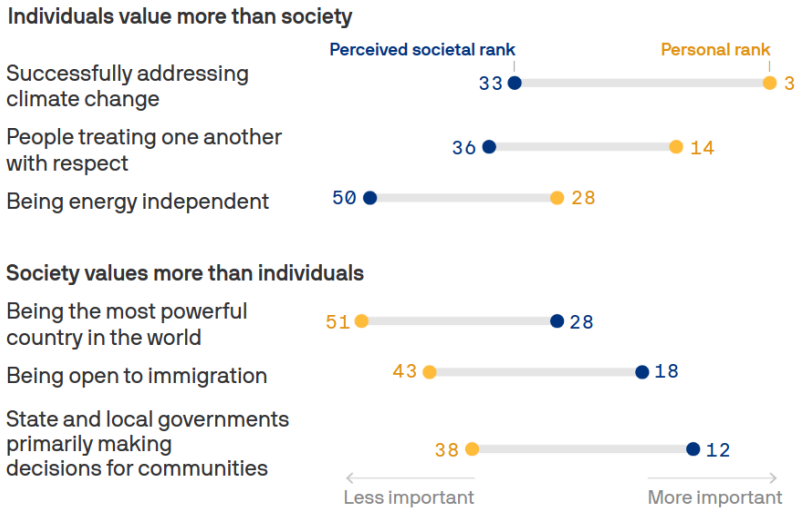
Despite the sharp divisions portrayed by media, most americans have far more agreement with respect to everyday policy issues

“Overall I think Americans want not to be divided as politics are forcing it to be, and that’s probably the biggest message of this poll,” said John Shattuck, director of the Carr Center’s project



Introduction

Despite the self-reported overlap in common values and ideologies, when asked how they believe 'most others' view the issue, it is often quite different.



Hypothesis

1.) Will a similar fixation on political figures or a small set of issues transpire in the respective subreddits?

2.) in the absence of tokens representing the above, will the model have a difficult time predicting the post origin?

Subreddits chosen

R/Conservative

R/democrats

This was soley based on participation, with the two seeming to represent the political parties presence on reddit

Pulling Data

10 requests, containing 10,000 posts each were gathered for each subreddit, totaling 200,000 posts.

Posts flagged as filtered by auto_moderator, moderator, or reddit were removed, which was the majority of the data.

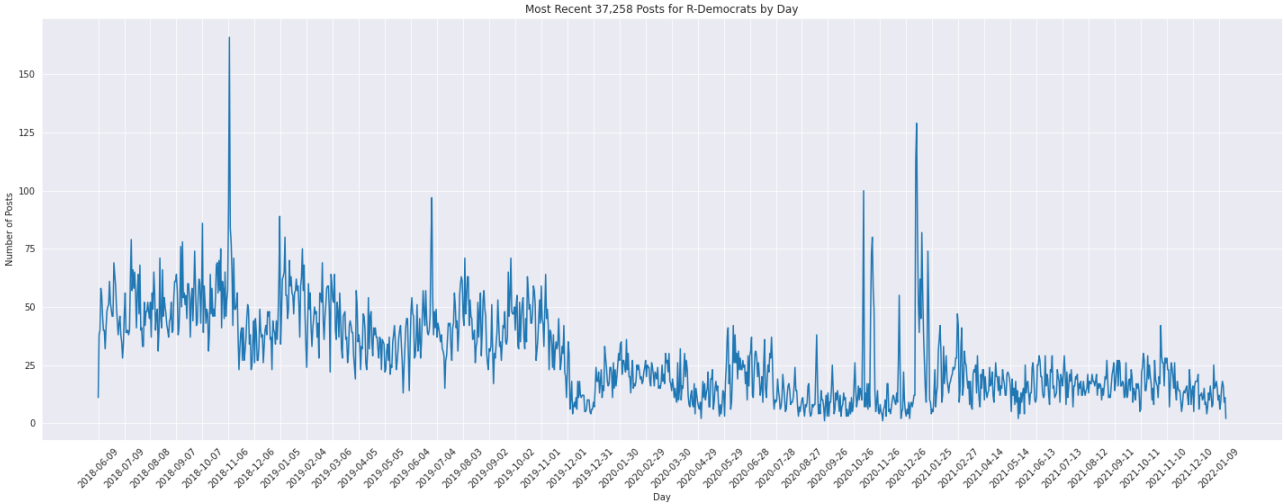
The final set consisted of 37,258 posts for r/democrats, and 35,662 posts for r/Conservative

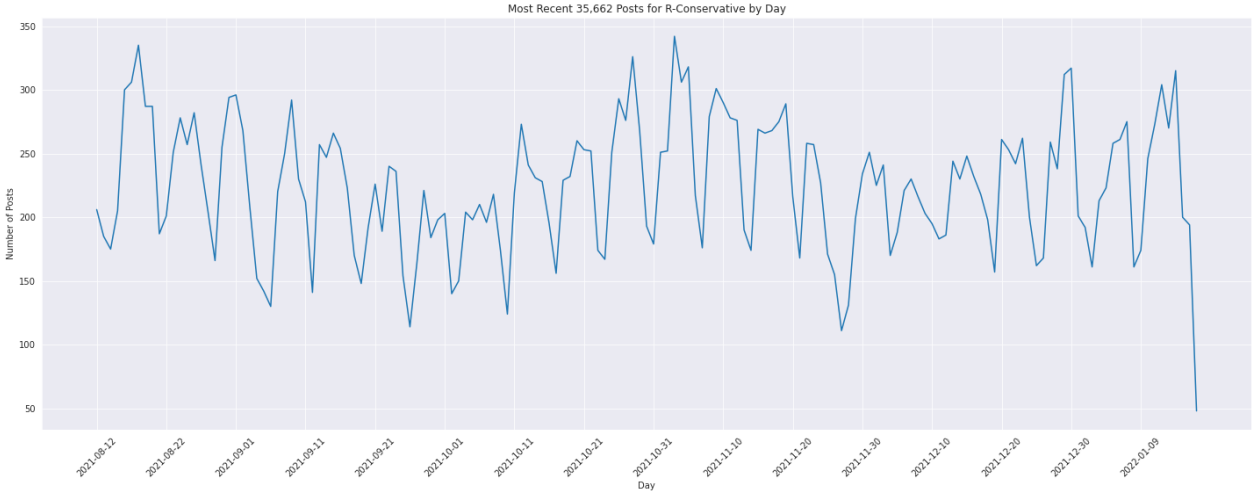
It became apparent these subreddits were highly targeted by bots

	removed_by_category
automod_filtered	31594
moderator	22073
reddit	10000
deleted	2057
author	4

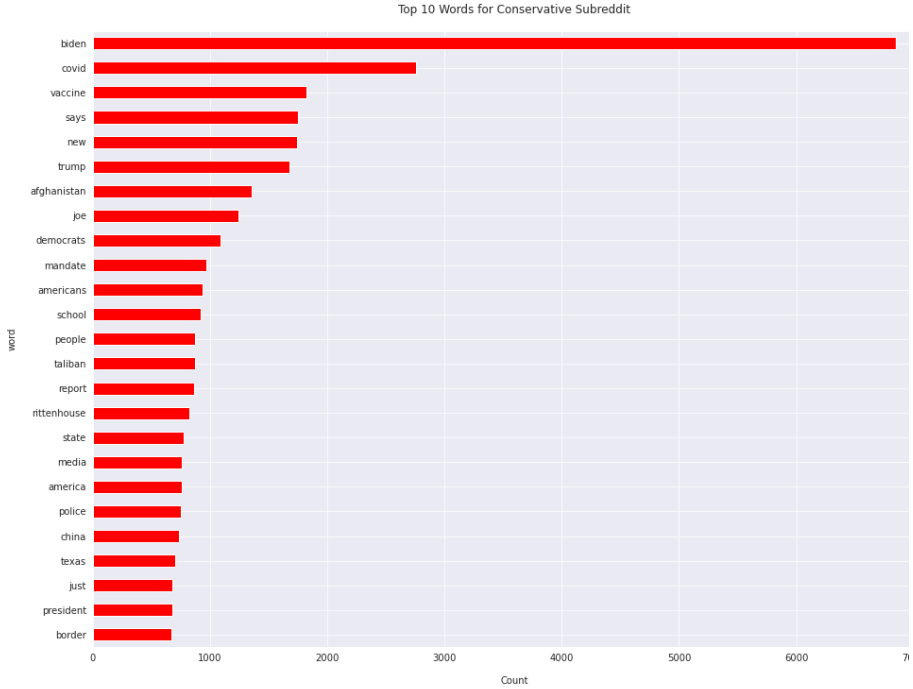
Differences in Activity

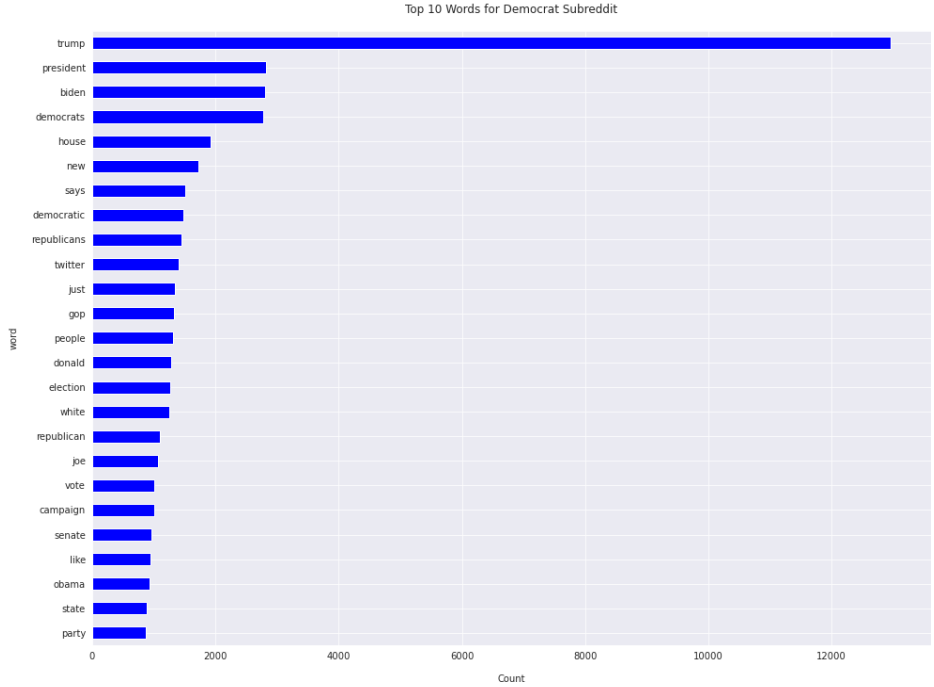
The r/conservative reddit had far more activity, as both had roughly 36,000, with the conservative subreddit having this many posts since August 2021, and the r/democrats going back to July 2018





Most Frequent Words in Titles





Model Results

Model	Transformer	Stop Words	N Grams	Max Features	Min_df	Max_df	train	test
Naiive Bayes	Count Vectorizer	.	.	4000	2	90%	0.831	0.819
Naiive Bayes	TFID Vectorizer	english	2	5000	2	.	0.838	0.821
Naiive Bayes	Count Vectorizer	.	.	10000	3	70%	0.852	0.828
Naiive Bayes	TFID Vectorizer	english	3	15000	.	.	0.867	0.837

Baseline

democrats 0.510943
conservative 0.489057

Limitations

A gridsearch using the Random Forest Classifier was attempted, some say its still running to this day

Though the 'max_df' variable accounts for words that appear in both subreddit posts, a more concentrated list of stop words could have brought to surface more discrete nuances

Limitations

Additional stop words, that I believed to be potentially useful were left in, and appeared in top 25 most frequent words

Author cross-posting not accounted for

Conclusions

Word choice appears to differ between conservative and democrat subreddits, as the model was able to predict the testing and training data with 85.2 and 82.8 percent accuracy respectively.

Since the activity was far greater for r/Conservative, the date ranges differed quite significantly, and with the pace of the news-cycle this could give much more context to this subreddit

Conclusions

More sophisticated classifiers such as the random forest, and extensive hyper-parameter grid-searches could likely fine tune the predictive ability, though requiring further processing time.

As discussed earlier, political engagement is higher among more far leaning ideologies, so with respect to the initial hypothesis of finding common ground between subreddits, perhaps this was not the right forum for such an investigation