An Analysis of Division in Political Ideology utilizing Reddit

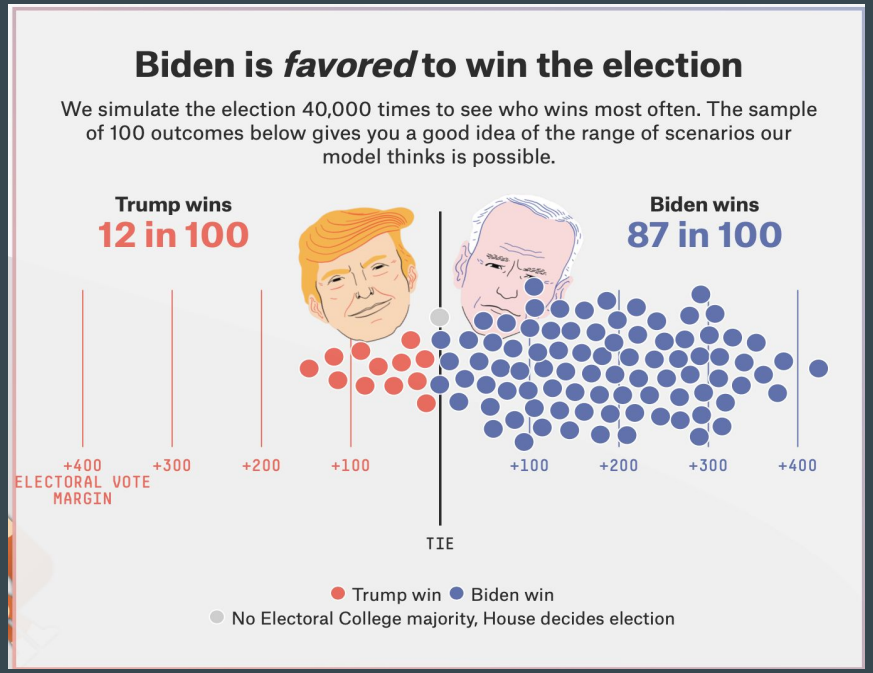
John Lawless

Problem Statement (Focus of Study)

- Client: a psychological research team in the beginning stages of data gathering. They are interested in assessing the level of division and polarization of Americans along the lines of the standard political spectrum.
- This will be done through analysis of two online political pages of differing political opinions on the online site reddit.com (online sources tend to encourage more divisive speech that may not manifest in face to face discourse).
- Through this, my client will attempt to address methods for assessing potential psychological risks inherent to this polarization, as well as possible methods for increased compromise.

Assumption - Polarization Leads to Entirely Different Data

From Five-thirty-eight and NY post



Assumption - Polarization Leads to Entirely Different Data

From Foreignpolicy and UsaToday: Two opinion

VOICE

Donald Trump Is Guilty

The only remaining question is what exactly he's guilty of.

BY MAX BOOT | DECEMBER 5, 2017, 3:58 PM

Pieces react to impeachment

Mueller report: Findings prove Donald Trump never colluded with Russia, obstructed justice

As far as collusion and subsequent allegations of obstruction by Donald Trump, there never was more to this would-be scandal than political innuendo.

Data Problem

- How many news sources do different political ideologies have in common?
- Will a sentiment analysis show a significantly different sentiment score between two different political groups?
- Is there evidence that different online political communities are talking about completely different topics?
- Can you build a predictive model that accurately identifies where a given post came from as evidence that this polarization is an actual feature of the two populations?

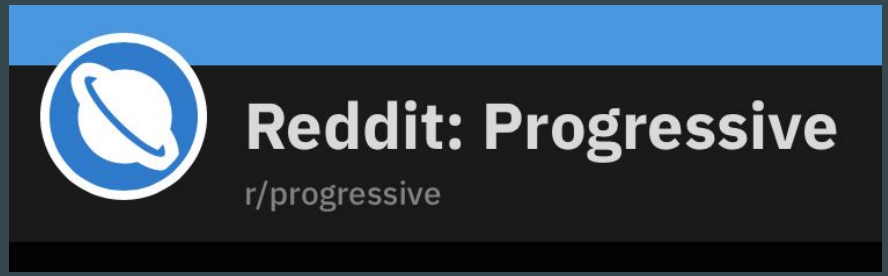
Data Acquisition

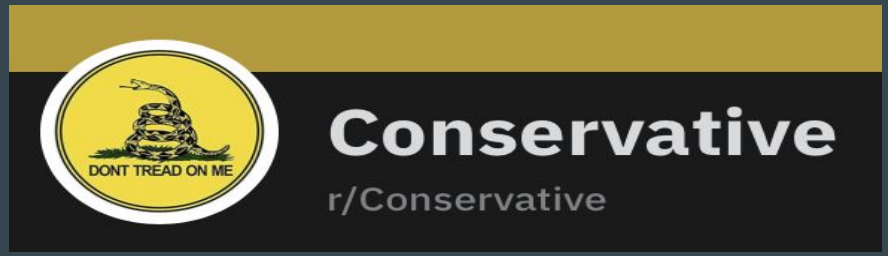
Data was acquired using The Python Reddit API Wrapper (Praw), using new, top rated, and controversial posts from two different subreddits.

Data was drawn repeatedly and then merged to get a substantial dataset for analysis.

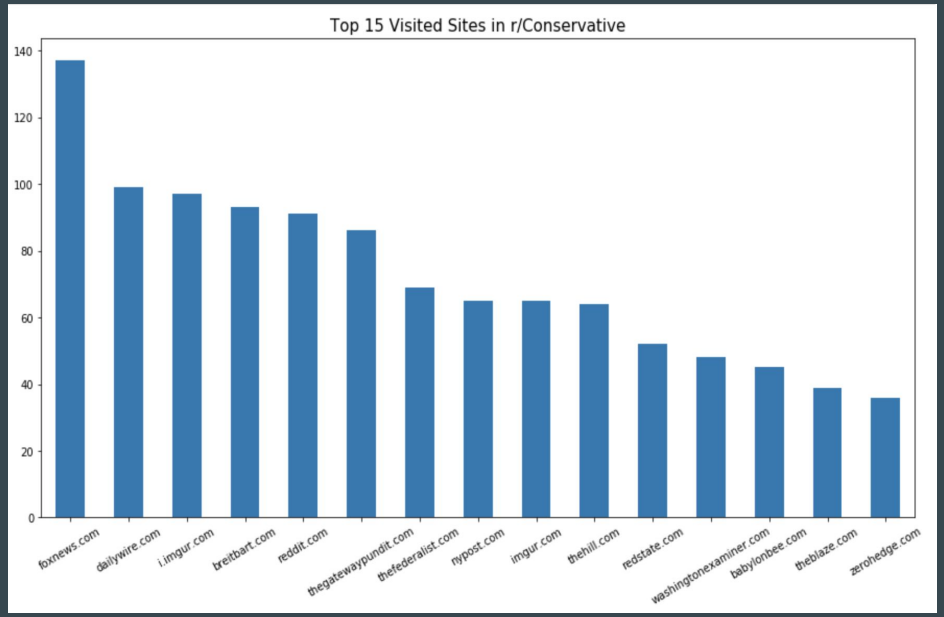
Initially, Conservative and Liberal were chosen. Progressive was more active, so it replaced Liberal.

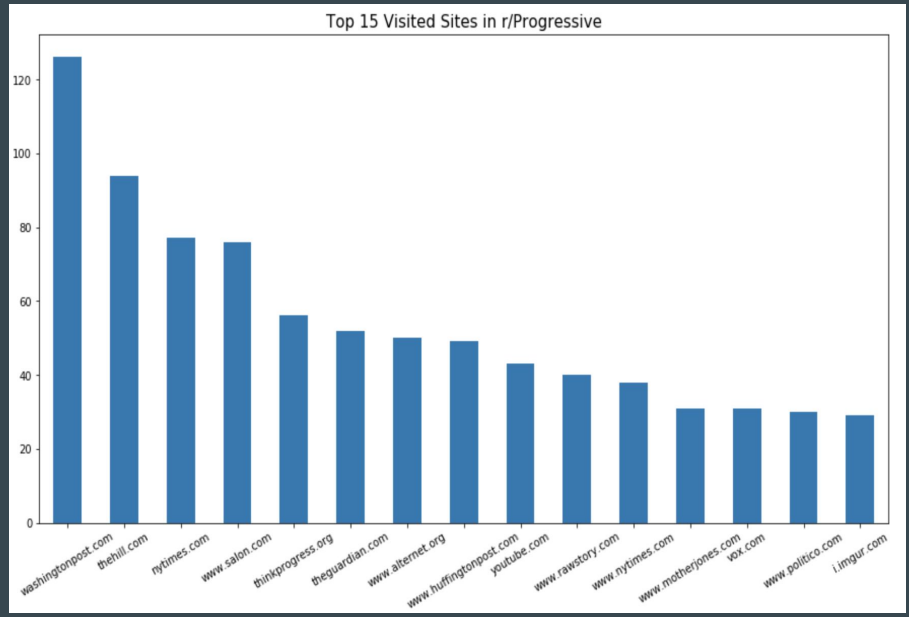
Using Praw's commentforest method (https://praw.readthedocs.io/en/latest/code_overview/other/commentforest.html) , comment data was drawn for analysis.





Where do Redditors Get their News?





Results - Where do Redditors get their news?

None of the top visited news sites are common among the subreddits

Only 122 (11%) of all 1116 news sources in data were common to both subreddits

This includes multiple sites such as imgur, reddit, youtube, etc, not news sites

This data is therefore strong evidence supporting the previous assumption of polarization among political groups.

Analysis of Sentiment on Text Data

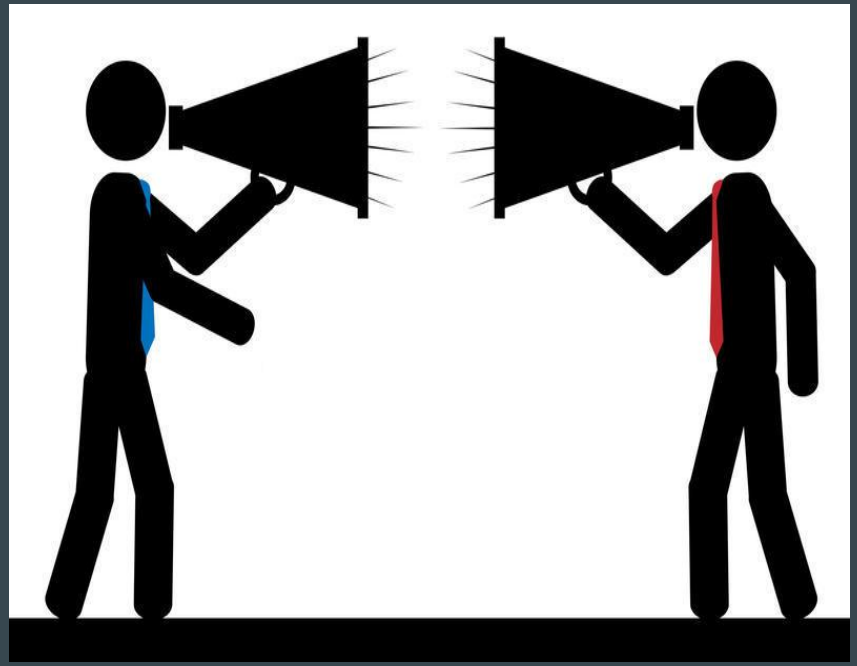
- All collected text data was analyzed using Vader SentimentIntensityAnalyzer
- The “compound” scores were problematic to interpret, so analysis was done on negative, positive and neutral scores.
- Due to more activity on Conservative subreddit, far more comments were extracted, leading to imbalance of data.

Results: Analysis of Sentiment on Text Data

subreddit	count	mean	std	min	25%	50%	75%	max
Conservative	2056.0	0.161911	0.039511	0.033	0.13575	0.159	0.185	0.357
Progressive	932.0	0.149964	0.048953	0.000	0.11500	0.143	0.179	0.426

The only score to show significant difference between the two communities was negative scores.

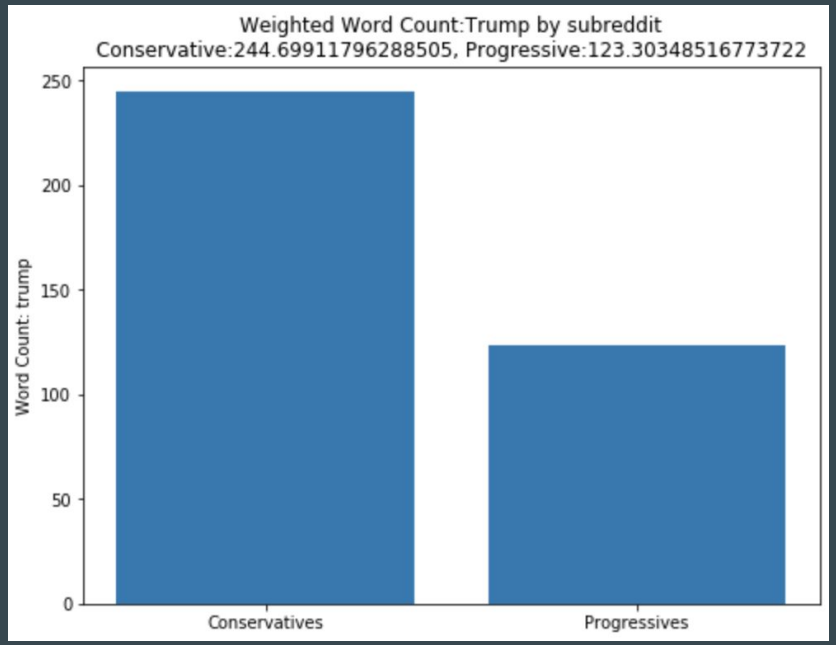
Analysis: Evidence of Polarization in online discussion

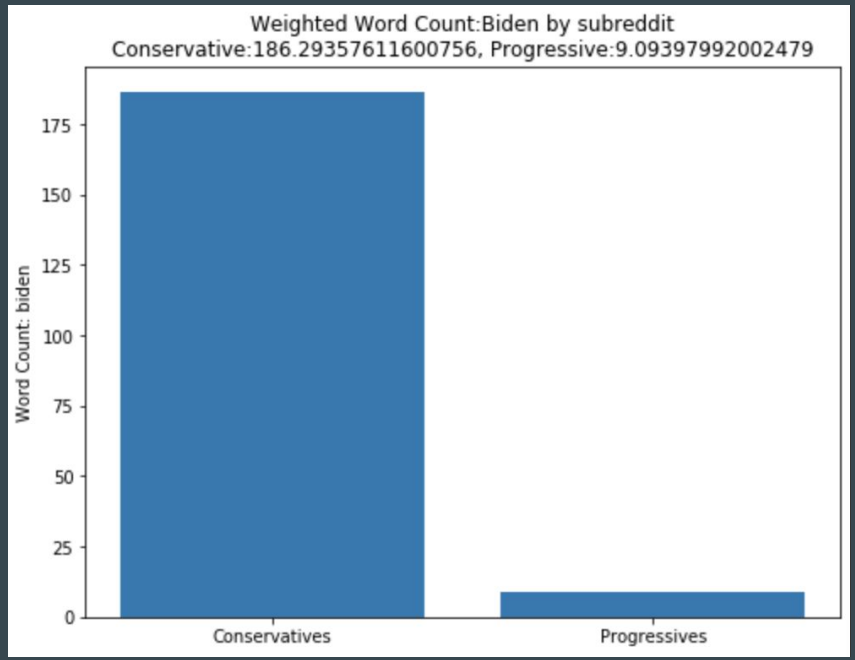


A TF-IDF count was utilized for this analysis, for two main reasons:

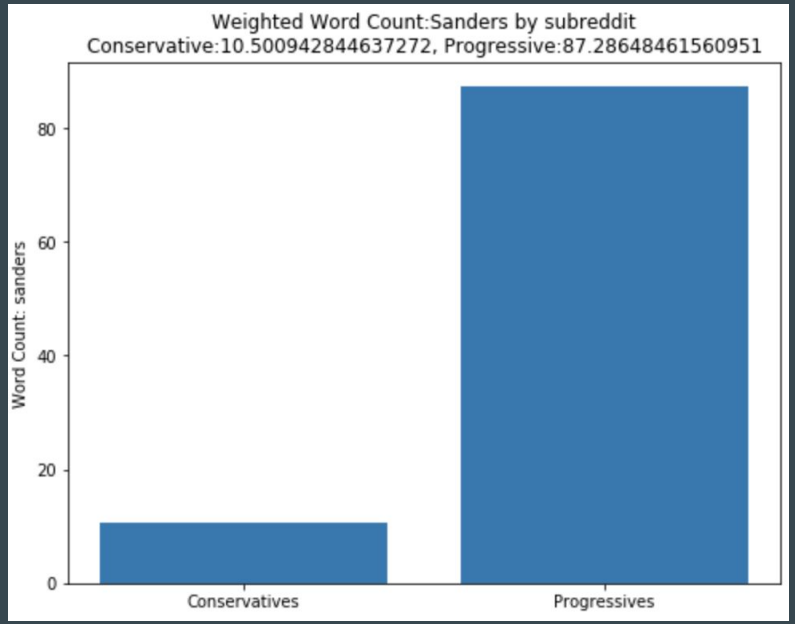
1) These subreddits have “topic” threads that center on singular discussion topics; heavily weighted words “matter” more
2) Subreddits are not balanced in activity; weighted scores is a method of mitigating an imbalance compared to pure word counts.

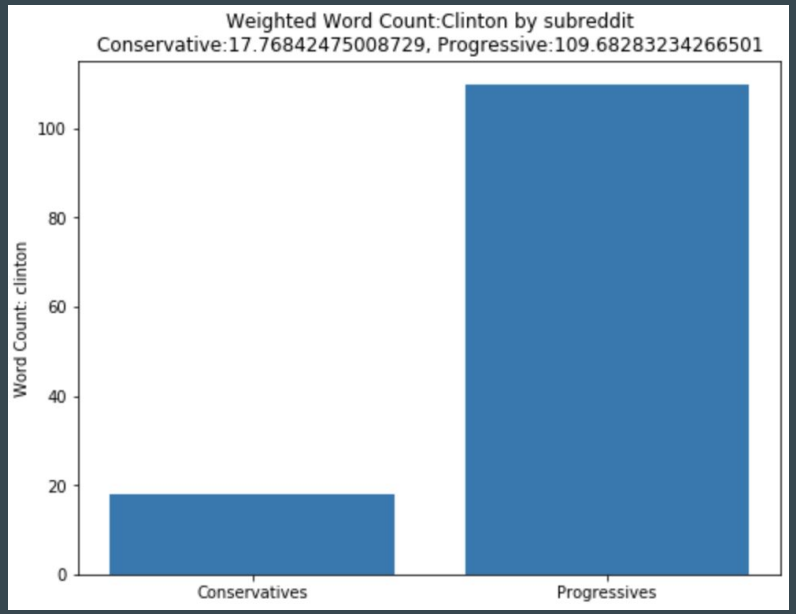
Words used more often in r/Conservative



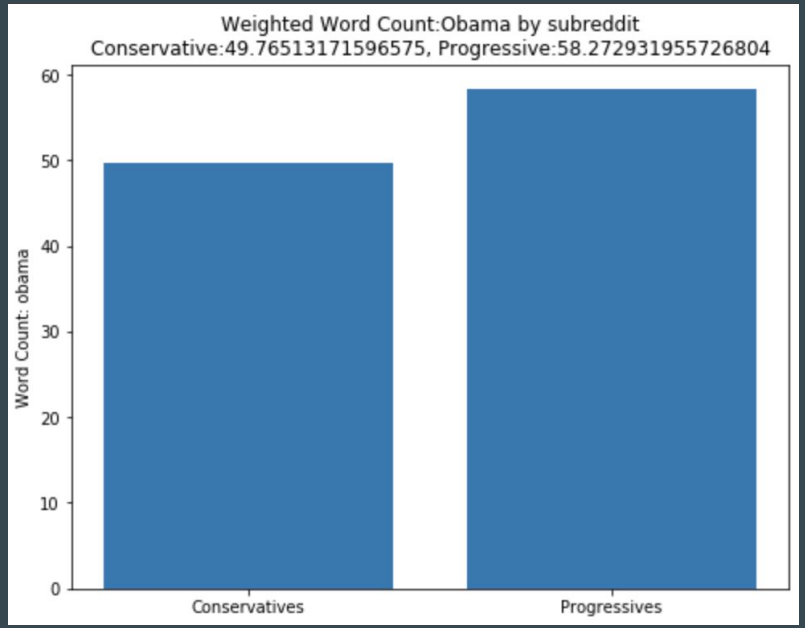


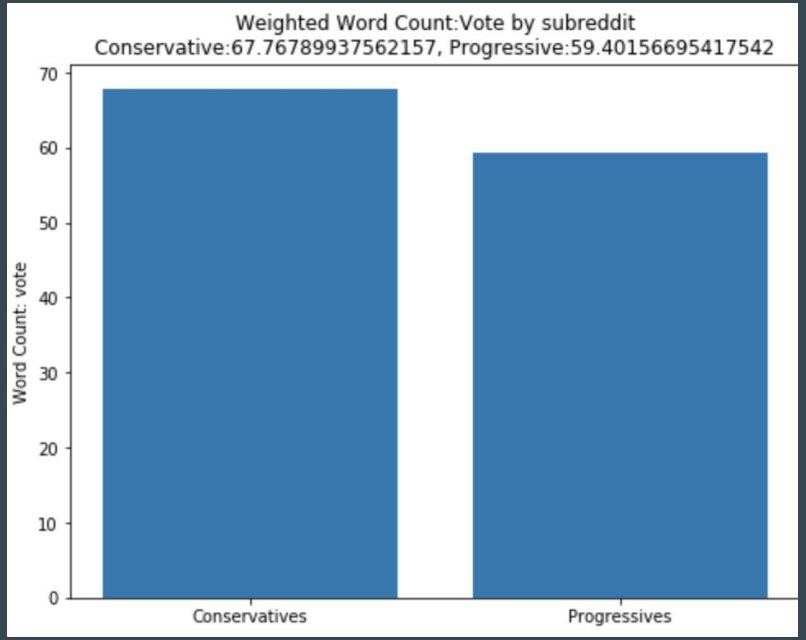
Words used more often in r/Progressive





Words found in common (fewer of these)





Results: Words as Evidence of Polarization

- The majority of words analyzed demonstrated a sharp contrast in score between both subreddits.
- As with news sources, this indicates further evidence of such polarization among political groups.

Modelling on the Data

- Three predictive models were used in an effort to accurately classify new reddit posts from each of these subreddits from texts/news sources
- Estimators: Random Forest Classifier, Support Vector Machine Model, and a Gridsearched Logistic Regression Model
- Due to data size and runtime constraints, other models were not gridsearched over
- Test data was drawn repeatedly, scored, then discarded

Modelling: Test Performance

Model results on unseen data :Test1	Accuracy on all features	Accuracy on words/news sites alone
Random Forest	89.9%	88.9%
Support Vector Machine	77.6%	77.6%
Logistic Regression Gridsearch (C:0.1, l1 penalty, lib linear solver)	89.6%	92.6%

Model results on unseen data:Test2	Accuracy on all features	Accuracy on words/news sites alone
Random Forest	88.6%	87.8%
Support Vector Machine	77.4%	77.4%
Logistic Regression Gridsearch (C:0.1, l1 penalty, lib linear solver)	91.8%	90.9%

Model results on unseen data:Test3	Accuracy on all features	Accuracy on words/news sites alone
Random Forest	87.8%	87.8%
Support Vector Machine	77.1%	77.2%
Logistic Regression Gridsearch (C:0.1, l1 penalty, lib linear solver)	91.1%	90.9%

Model results on unseen data:Test4	Accuracy on all features	Accuracy on words/news sites alone
Random Forest	87.6%	87.1%
Support Vector Machine	72.0%	71.9%
Logistic Regression Gridsearch (C:0.1, l1 penalty, lib linear solver)	91.3%	91.3%

Model Performance

- Baseline - 50% accuracy
- SVM had highest bias and variance of all models
- Random Forest Classifier consistently scores accuracy of 87.1-88.9% on unseen words and news site data
- Linear Regression model scores same data between 90.9-92.6%, showing least bias and variance of all models

Results

- Through both analysis and predictive modeling, this study is evidence of strong political polarization in online communities.
- These subreddits are reading completely different sources, and having entirely different conversations.
- This data can serve to get my client started on the process of identifying potential risk factors among those who are strongly politically polarized on both ends of the spectrum.

Resources

reddit.com/r/Conservative
reddit.com/r/Progressive

https://praw.readthedocs.io/en/latest/code_overview/other/commentforest.html