

Introduction

In this project we mine data from a public database of Reddit comments and use it to observe and understand better the echo chamber effect. We perform an observational study on the interaction between Reddit communities centered around political discussion. More precisely, we are interested in creating and demonstrating techniques to measure such phenomena. We introduce different approaches to that problem. One of our objectives is to raise awareness of how a big data operation (like a political campaign) could leverage the existence of such echo chambers in mass media.

The Echo Chamber Effect

We say a community is an echo chamber when it only interacts relatively within itself. Similar users are brought together (homophily) and then they influence each other, reinforcing the traits, biases and opinions they share. When a community or a group of communities is established, this then creates a self-sorting effect: each user finds their comfort zone, and isolates themselves from the outside. This is detrimental in the context of healthy debate and a major contributor to opposing views polarizing and becoming extremes, and the reason we want to be able to measure it.

Reddit and Karma

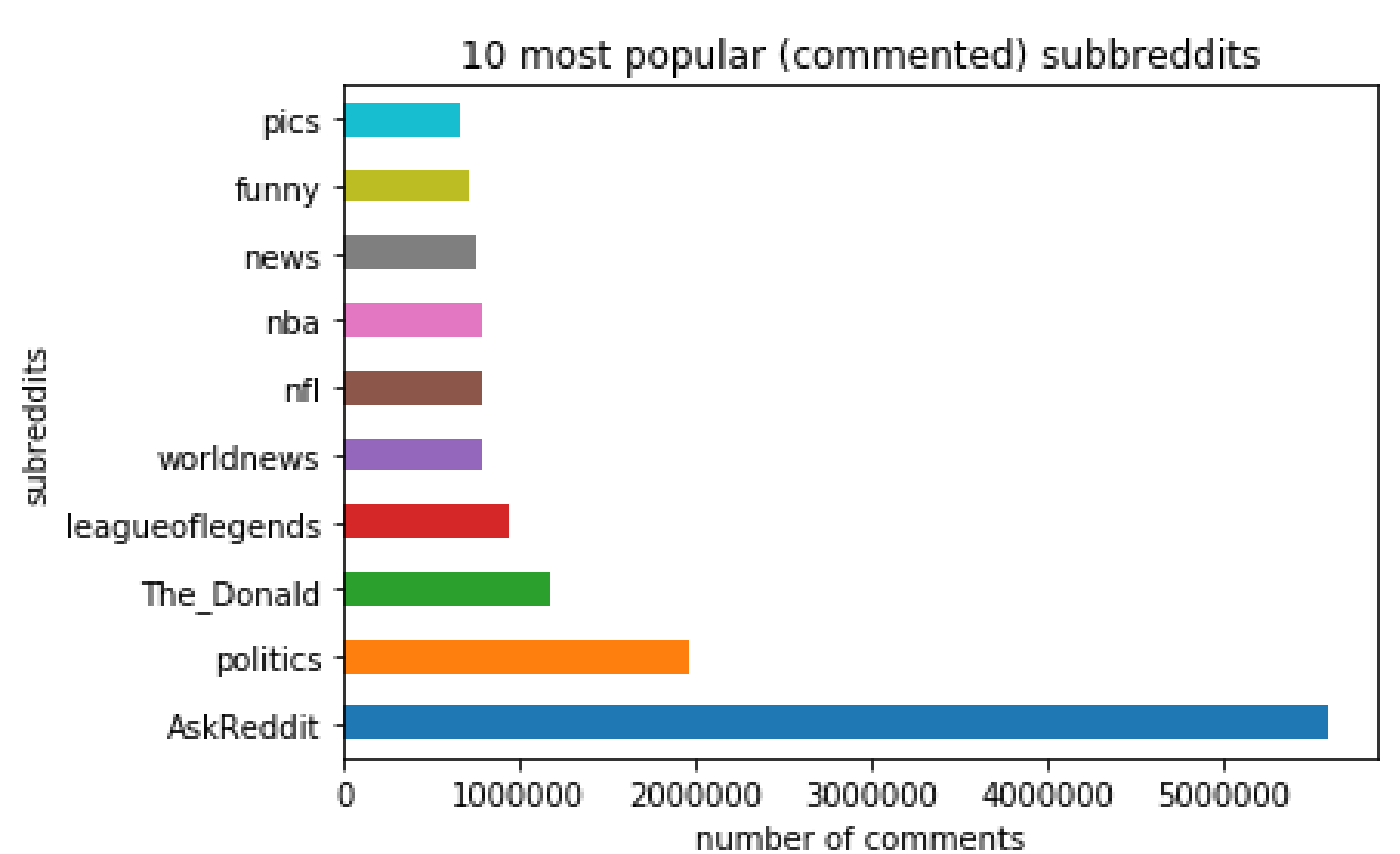
The hallmark of Reddit is the karma system. Upvotes and downvotes from users determine what information other users will see and thus become relevant. A known flaw of majority voting is that it is not the most efficient system at taking into account diversity of opinion. We use votes as one of the key features in both our approaches to measuring the echo chamber effect.



Data Mining

For this project we had access to all of Reddit's comments since its creation in 2005 until early 2017. Since we focus on political discussion we decided to go with a fitting subset, including the year of the US election, a year where discussing politics was commonplace due to the Trump campaign. Here's the preprocessing our big database underwent to become tractable:

1. Retrieve data from 2016
2. Select desired features: content, author, score, controversiality, subreddit, etc
3. Filter for the top 200 subreddits with the most comments
4. Eliminate outliers: r/AskReddit and r/counting
5. Remove [deleted] and [removed] comments and authors
6. Subsampling: randomly select ~1M users that commented in a political sub at least once, take all of their comments, ~64M comments total
7. Focus on politics by identifying the 8 political discussion centered communities in the top 200: r/politics, r/The_Donald, r/SandersForPresident, r/HillaryClinton, r/Conspiracy, r/EnoughTrumpSpam, r/politicalDiscussion and r/UKpolitics



Local Approach

We tried to model how polarized the communities are in terms of political lean by means of analyzing the content of the comments. We perform topic modeling and other NLP techniques to see if we can observe clear differences in the composition of each community and if really they are segregated by ideology. If communities have a tendency to talk in a certain manner or prefer certain topics strongly then the network will self-sort individuals with regards to what their biases are. That is one indicator for the presence of deep echo chambers.

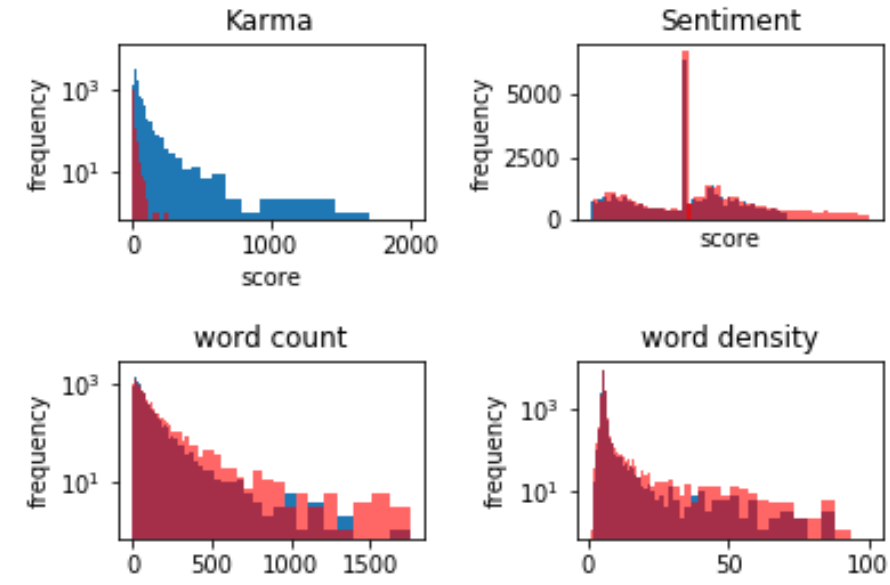
Natural Language Processing

In this approach we make use of as much features and metadata as possible to create a classifier to distinguish users in terms of /left/right political lean. As training set we took comments from 4 explicitly political communities: r/Republican and r/Conservative for right-leaning comments, r/democrats and r/Liberal for left-leaning comments (at least in the American sense).

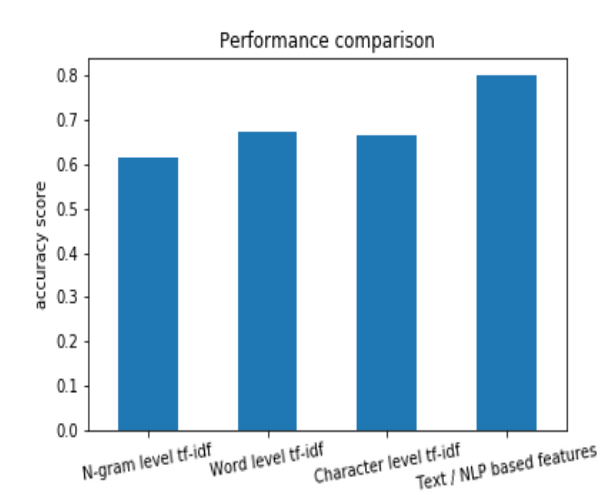
Feature Engineering

As a first step, we performed sentiment analysis on the content of every comment in both our training set and the working samples. Other than sentiment, we calculated a variety of metadata for each comment:

- Character count
- Word count
- Word density
- Punctuation Count
- Title Word count
- Uppercase word count
- Noun count
- Verb count
- Adjective count
- Adverb count
- Pronoun count



Algorithm Choice

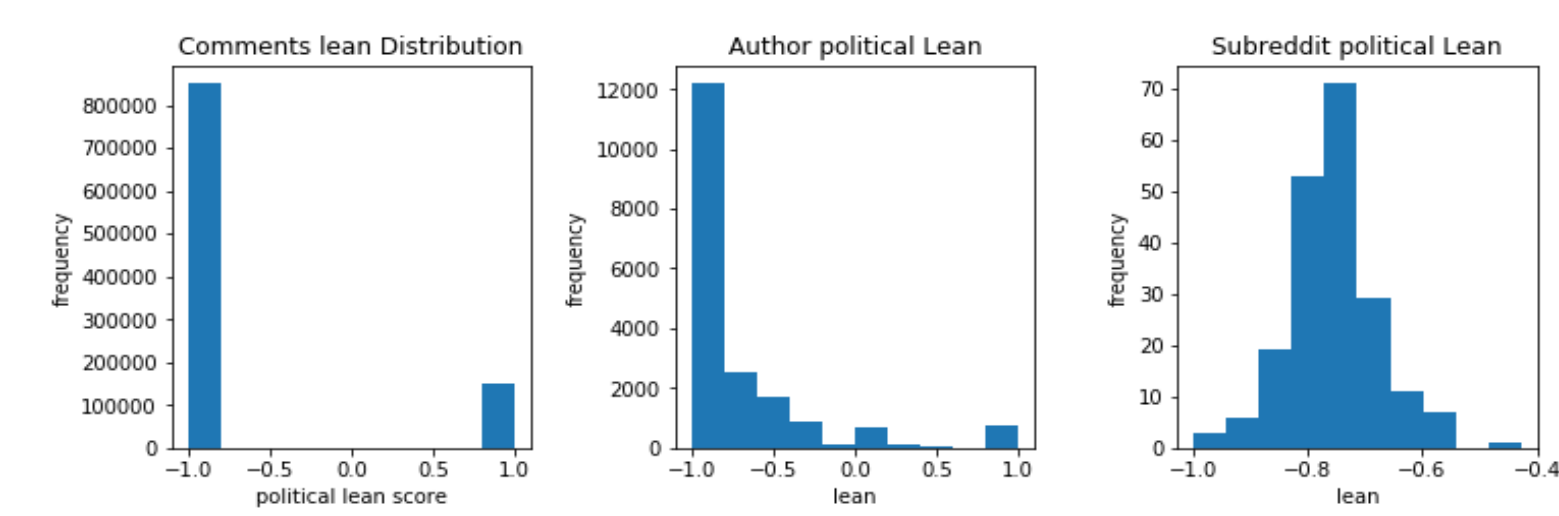


For the classification task several stock methods for text representation were tested. The best score in terms of accuracy (correctly classified the most comments) was selected

- N-gram level TF-IDF
- Word level TF-IDF
- Character level TF-IDF
- Text/metadata based features

Results and Interpretation

According to our best classifier and given the assumption that if a comment "looks" like it belongs on the right/left subs then it must be respectively right/left lenient, we gave political lean scores at 3 levels: comment, author and subreddit. Each author's lean is the average of their comments and a subreddit's lean is the average of the authors that commented on it.



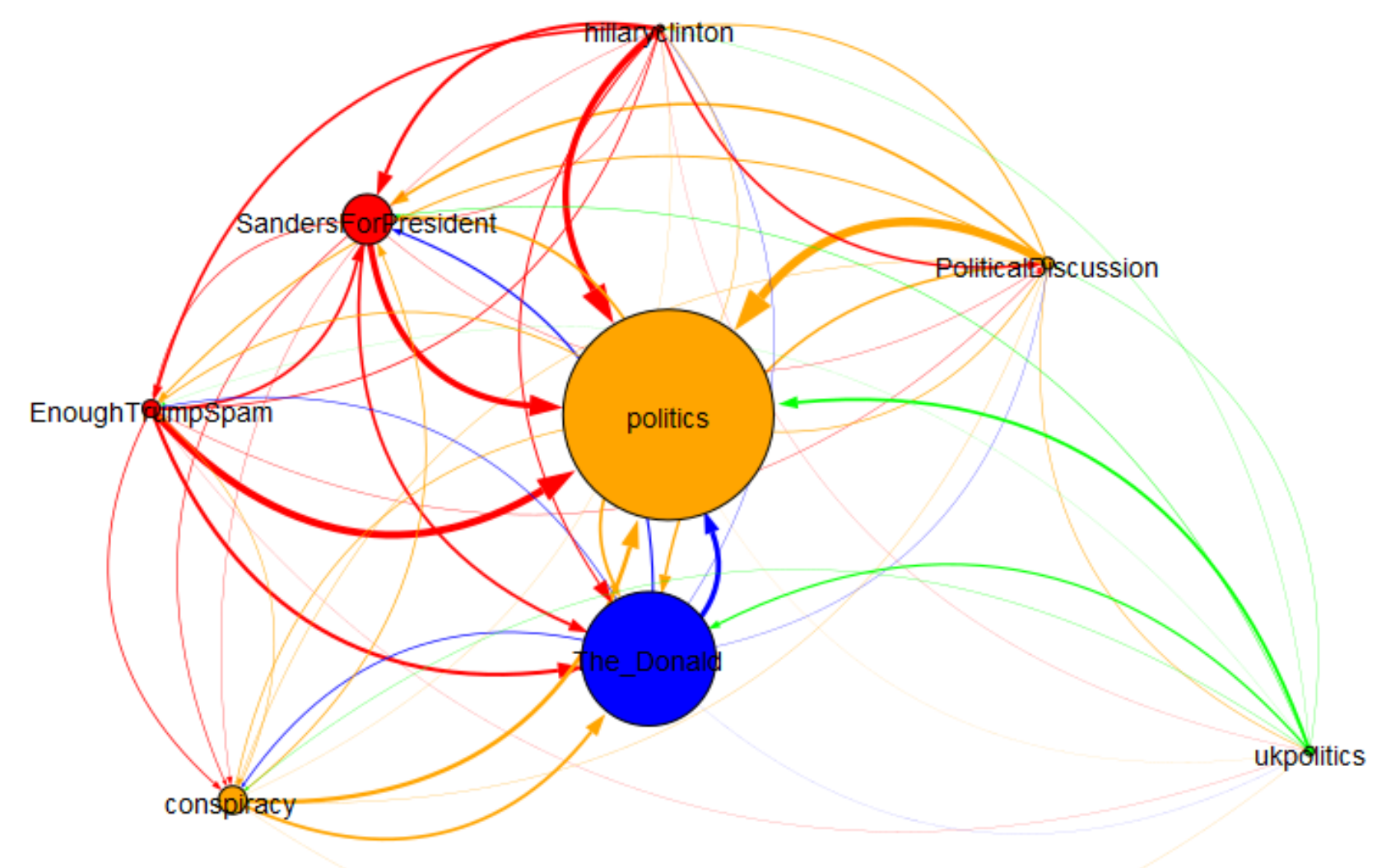
As we can see on our final classifications it appears as if most of the comments and authors are very left-lenient. This makes sense, given that the left-leaning communities are the biggest in terms of users. So we can infer that Reddit as a whole is a mainly democrat aligned hub. This fact talks about the inherent bias of the network as a whole. Two points can be made: for one, left-leaning ideas are going to be dominant (you could say that reddit is a liberal echo chamber) and right-leaning ideas are going to be treated as dissent (right-leaning users are going to remain within their isolated groups). This relates to our global analysis.

Global Approach

Looking at the problem from the perspective of interaction between communities, we are interested in spotting groups that interact less than average with outsiders. With that in mind we formulated a model using network graphs and used heuristics to hopefully capture exactly that. If a community is more isolated it is more likely to be considered a candidate for the title of echo chamber.

Network Models

Now we don't take into account the actual content of the comments like we previously did. Instead, we keep it simple and just look for activity: for each user, assume they are ideologically aligned with the subreddit they successfully (aka get upvoted) post the most on. Now that each user has a "home" community we can observe how that community interacts with another by drawing a thicker directed edge the bigger the proportion of users in that community that also comment on the other.



When applying our model to the 8 subreddits that revolve around politics among the top 200 subreddits this is what we see. r/politics is the main hub for political discussion and is at the center of the network. Surrounding it there are several other communities that are more aligned with certain ideas/topics. As we can see, r/The_Donald, a community centered around praising Donald Trump shows almost no interaction with the rest of Reddit. The users that comment in that sub seem to not be exposed to ideas on the outside (in fact, they're pretty much banned by their rules). This is consistent with our previous analysis: right-leaning individuals are left to isolate in communities on their own. We found it! A **prime example of echo chamber** in the wild!

Finally, centered in politics as we are, we retrieved data from communities dedicated to specific political movements. The findings are interesting: as we can see, in terms of number of interactions, ideologies that are closer in the political spectrum appear closer together and interact more.

