Community crossover: classifying comments from political Reddit
by Fauna Mahootian

Abstract:
Communities and politics are becoming increasingly polarized. I endeavor to examine collective opinion in Reddit communities through comment scores. I do this by predicting how a comment in one reddit community (for instance the Donald Trump community) would fare in another (for instance the Hillary Clinton community). Would it be well received, or poorly received? I created Naive Bayes classifiers to approximate this judgement. The classifiers work well for predicting positively scoring comments, but poorly for predicting negatively scoring comments. I then used topic modeling to see if I could get an idea of subject matter talked about in positive and negative comments for a given community, to explore any differences. The topic models were not very elucidating beyond providing a list of topics for all of the comments. I then discuss possible confounding factors that give insight on how to improve this project in the future.

Introduction

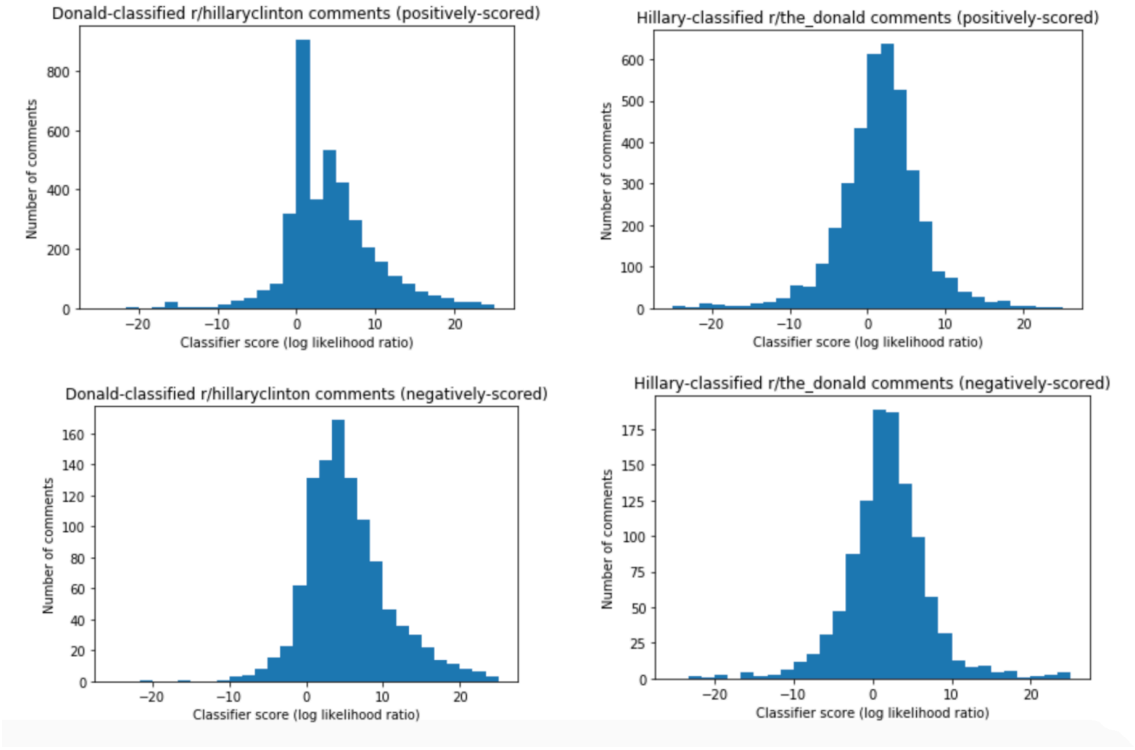
I'm very interested in the formation and dynamics of communities. I recently read a 538 article[1], published shortly after the 2016 election, in which the author characterizes Reddit communities (subreddits) in terms of each other based on member overlap. They looked in particular at r/hillaryclinton, r/SandersForPresident, and r/the_donald, the primary subreddits for supporters of Hillary Clinton, Bernie Sanders, and Donald Trump, respectively. I found this article very interesting and I enjoyed their method of characterizing communities (by using the network of commenters and seeing what communities overlapped with the given community, on this basis). I wanted to do something similarly political Reddit community themed. I was interested in how comments typical of one subreddit would fare in another. Comment score is a proxy for the community's opinion of a comment; comments can be upvoted or downvoted, and the sum of these two measurements results in the comment's score. I hypothesized that negatively scored comments in r/the_donald might be favorably scored in r/hillaryclinton, and vice versa.

Classifiers

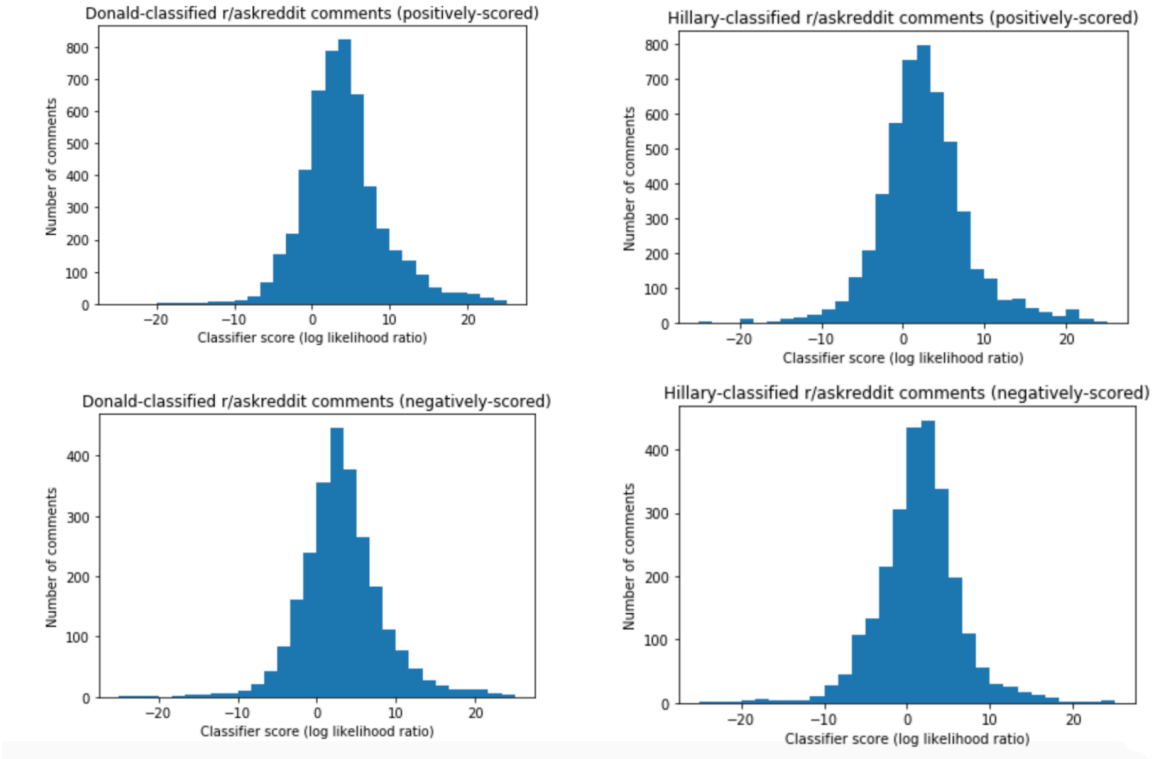
I chose to create comment classifiers based on comment score for r/hillaryclinton and r/the_donald. I used comments from the aftermath of the 2016 election (with dates ranging from November 2016 to May 2017). I gave the Hillary classifier comments from r/the_donald, and the Donald classifier comments from r/hillaryclinton. These classifiers are Naive Bayes classifiers, which essentially classify a given set of words as one of two categories by comparing the probability that it would occur in either category. In this case, if the rating is negative, then the classifier predicts that the comment would be more likely to have a negative score than a positive one, in the subreddit that the classifier was built from. If positive, it predicts that a positive score is more likely. The higher (or lower) the number, the more probable the classifier thinks it is that the comment belongs in the given category.

---

[1] https://fivethirtyeight.com/features/dissecting-trumps-most-rabid-online-following/



I also classified some comments from a third subreddit, r/askreddit, for comparison. This reddit talks about all sorts of things and from what I could tell doesn't have a political focus, which is why I chose it for comparison:



As you can see, the Hillary classifier classifies r/the_donald comments the same way

regardless of whether they are positive or negative. The Donal classifier classifies r/hillaryclinton comments a little more discriminately, although the majority of its ratings are positive. When validating my classifiers, I found that they were pretty accurate at classifying positively-scoring comments (about 80% accuracy), but not very good at classifying negatively scoring comments (about 30% accuracy). It is clear in all the above graphs that the classifiers have a tendency towards classifying positive.

Topic model

I made a topic model in order to get a better idea of how communities score their tweets, to see if the topic model could expose or explain any patterns there. A topic model is essentially a group of topics that could be used to describe a body of text. Each document in the text body is composed of different percentages of all of these topics. I used the positive and negative comments from all three communities I've introduced to make the model.

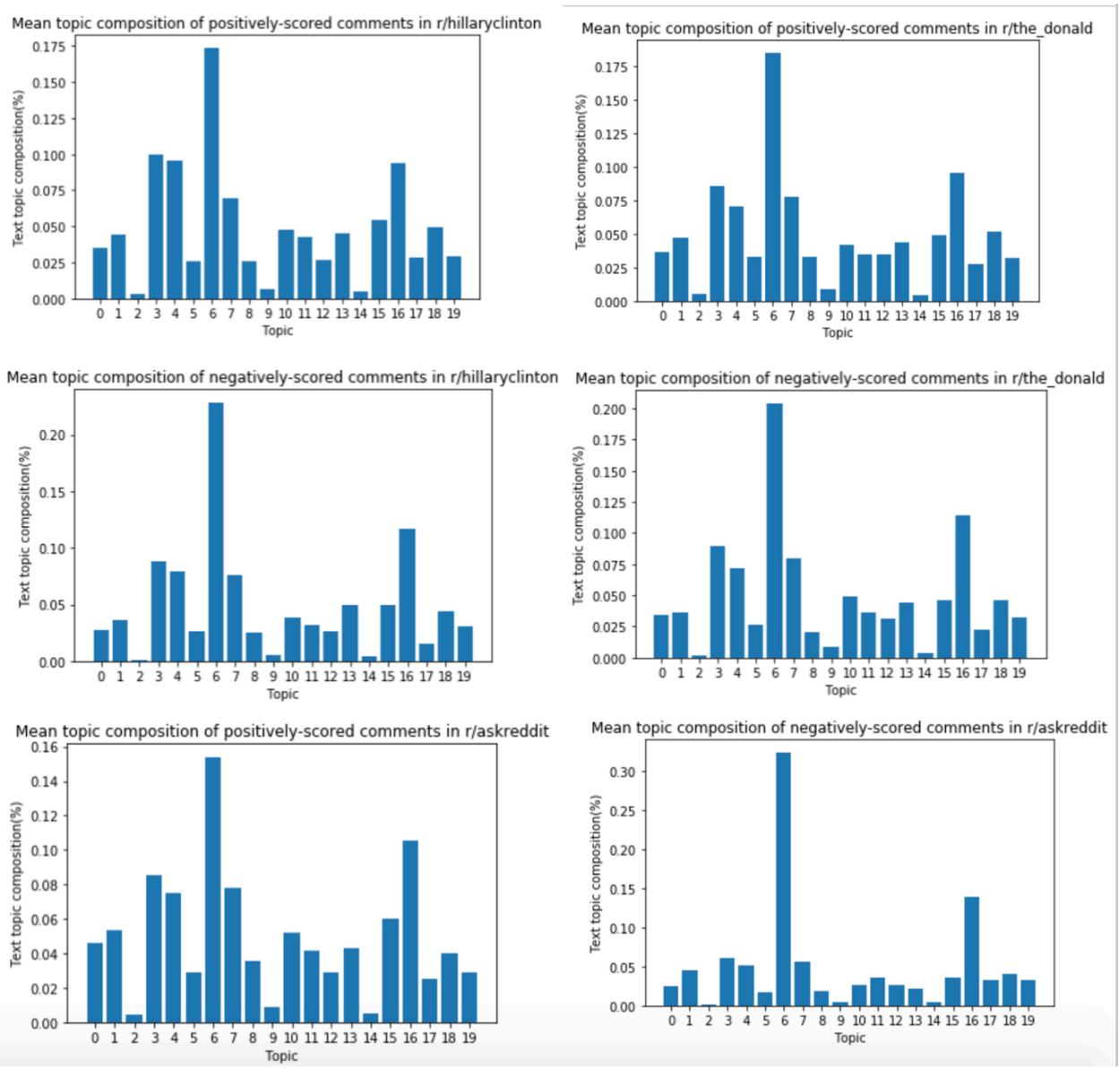
These are the topics my model settled upon (I've bolded the ones I thought were relevant to r/the_donald and r/hillary). The topics are comprised of all the words that were labeled as that topic, in descending order of frequency. These are the top 10 for each topic.

0. car gun stop driving called cars light drive speed road
1. show season episode character movie watching watch good watched series
2. abc schiff httpstwittercomconflictsstatus httpwwwgodlikeproductionscomforum message russian ukraines donbass ties males
3. problem people system reason society good point change part fact
4. money work pay people make tax buy government dont company
5. trump comey president news fbi russia hes media investigation hillary
6. people dont youre good doesnt make thing things point person
7. food eat drink hair pizza hot good water beer restaurant
8. burn onions dinosaur weight big dinosaurs eat ocean fat animal
9. vote pen house state macron senators call district french round
10. school college high year kids students university teacher class schools
11. women men love sex rape brick make sexual god centipede
12. trump hillary bernie party vote clinton election people democrats voters
13. god religion science climate islam muslims jesus christian world bible
14. questions post automatically removed action contact concerns question bot performed
15. war country world american america countries government france nation germany
16. time day back didnt years guy told friend kids friends
17. music comments tag child parent posts notice play usa song
18. post news read facebook comment media reddit edit video article
19. white people black racist women hate theyre rights race men

On the following page are the graphs of the topic composition of the set of positive/negative comments for each subreddit. As you can see, there is not a dramatic difference between r/the_donald and r/hillary, which is expected, as they both talk about partisan issues. There is however a difference between these two and r/askreddit, since the topics discussed in that sub are much more broad.

Because there's not much of a difference between the topics for positive and negative comments within a community, I'd argue that the topics aren't the sole determinant of a comment's score. There are two missing pieces of information that could be relevant to comment score that topic models can't explain: word orders as a determinant of intent/meaning, and the context of the comments in whatever discussion was happening. Perhaps downvoting occurred because of the context plus the comment content, rather than solely the content. Looking at many of these comments myself, I can't figure out why some of them were downvoted. Another issue that might trip up the topic model (and classifier) is that Reddit

comments can be removed if they violate community guidelines, or by the original poster. If this feature didn't exist, those comments may have been downvoted instead. Or maybe they were downvoted before they were removed. I'd wager that some of the language use in /r/the_donald would get a comment removed in /r/hillaryclinton. But that "negative rating" data is missing because those comments have been removed. So my classifier handles slurs or negatively signifying language it doesn't have in its vocabulary as neutral.



I created classifiers to see if the way that each community scores their own comments is predictive of how they'd vote on comments made in communities outside of their own. The classifiers behaved slightly differently from each other. Next, I used topic modeling to see if I could get an idea of subject matter talked about in positive and negative comments for a given community — to explore if that might be a factor that differentiates them from each other. Before implementing this project, I hypothesized that negatively scored comments in r/the_donald might be classified positively in r/hillaryclinton, and vice versa. However, there are

many potential confounding factors to my causal assumptions behind my hypothesis, which I will now explain. This will also help to elucidate future directions to take this work if I want to continue with it.

Discussion: potential problems with my approach:
My biggest failing was the difference in number of positive and negative comments for each subreddit. I had something like 4000 positive comments and 1000 negative comments. This difference definitely affects the accuracy of the classifier.

Many of these comments have scores of -1 or -2. This means that they were probably downvoted by the person who the commenter was arguing against, making the downvote not the view of the community, but of one or two people. I think I should have picked a threshold for net score. If the post has many more downvotes than upvotes, that seems like the community's belief. Although I'm not sure what percentage of the community typically participates in discourse on a given post. I think a score of -6 or above seems like a good starting point.

Most of reddit is liberal, so my choice of r/askreddit as a "neutral" comparison isn't really a good control. However, Clinton's politics are pretty center, so I'd say Reddit is at odds with her as well, but less so than Trump.

I collected positive and negative comments between different time scales. It doesn't matter across communities, but it does matter within communities; it changes the prior likelihood of a comment to be positive or negative. Prior likelihood of a comment being positive is number of positive comments divided by total number of comments collected. I should've collected the comments within the same time period. For example, for r/hillaryclinton I collected all the negative comments between Nov 29 2016 and May 14 2017, and all the positive comments from May 1 2017 to May 14 2017. The resulting prior probability of negative comments was much higher than it should have been, which throws off the classifier's ability to predict accurately. Although if I had sampled more accurately, the prior likelihood of positive comments would be way higher.

The following section contains a more scientifically rigorous description of my methods.

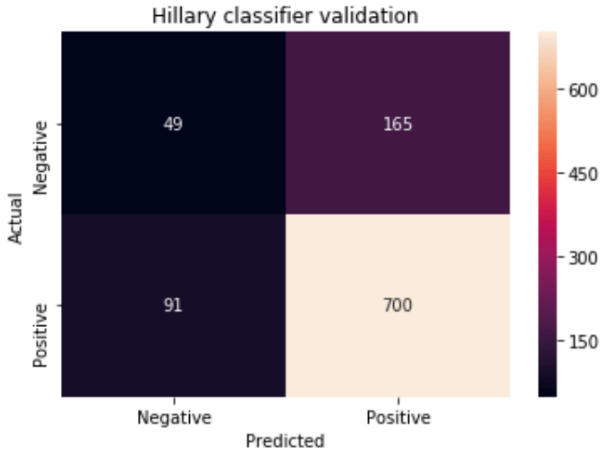
Methods

To gather my data, I used the pushshift.io API to query Reddit. I restricted the date of comment creation to be before May 14th 2017 (an arbitrary date shortly after the election). I made two collections of comments for each subreddit: ones with positive scores and ones with negative scores. I queried from three subreddits: r/hillaryclinton, r/the_donald, and r/askreddit. I chose askreddit as a "neutral" basis for comparison because it seemed like it would have subject matter all across the map and doesn't seem like it would attract partisanship (like for instance r/guns). After acquiring the data, I filtered for comments that were longer than 20 words and removed punctuation so that I could generate a topic model. I ended up with: from r/ hillaryclinton, 1071 negative comments from between Nov 29 2016 and May 14 2017, and 3909 positive comments from between April 19 2017 and May 14 2017; from r/the_donald, 1082 negative comments from between May 1 2017 and May 14 2017, and 3833 positive comments from between May 14 2017 and May 14 2017. I reserved 20% of the comments in each category for testing. The rest I used to train the two classifiers.

I chose to implement a Naive Bayes classifier because I believed that the independence assumption held for my dataset[2]. I validated my classifiers using cross-validation: they

[2] https://nlp.stanford.edu/IR-book/html/htmledition/properties-of-naive-bayes-1.html

perform pretty well when given positive comments — about 80% accuracy (according to F1 score) (83% for Donald and 85% for Hillary). But they perform terribly on negative comments — about 30% accuracy (29% for Donald, 28% for Hillary). (see below model for accuracy heatmap). I suspect that this is because I had so many less negative comments than positive ones to train and test with.



I used Mallet to create an LDA topic model to get more information about the content of the comments. I used 20 topics, and removed stopwords before making the model. I didn't use the topic model to generate quantitative information, but as a way to look for trends.