Spring Project Proposal

Faiz Surani, Rahul Araza, Victor Bai

April 12, 2019

1 Introduction

Political polarization in the United States has vastly increased in the past several decades [1]. Over this time period, Americans have become more likely to identify as liberal or conservative rather than moderate and less likely to bear positions from both sides of the political spectrum [2]. Political scientists have cited several causes for this trend including a growing generational gap, a transforming economy, and the gradual expulsion of moderate politicians from both parties. One factor, however, has caused the most concern among researchers in recent years: the rise of social media. While social media has connected the world in ways difficult to imagine just two decades ago, it has also facilitated the increasing polarization of political opinion due to the nature of social networks to form bubbles of likeminded individuals in online groups at the expense of a diversity of opinions [3].

Social media's polarization effect cannot be solely explained by its tendency to form echo chambers, however. Research has shown that exposure to opinions of an opposing ideology on social media over a period of time can in fact *increase* political polarization [4]. Some have postulated that political discourse on social media between opposing sides has seen a significant decline in civility and quality over the past decade, especially in the wake of Donald Trump's rise to popularity and subsequent election to the presidency in 2016 [5]. Especially on websites that host many diverse political communities, this effect has been compounded by bad-faith participation in others' discussions in the interest of derailing discourse, a practice known as "trolling" [6]. One such website is Reddit, a social media website founded in 2005 that has come to host vastly diverse user-run communities centered around common interests[1] known as "subreddits" [7].

Reddit hosts hundreds of subreddits dedicated to political discussion and activism all across the political spectrum. The general tendency of Reddit users to filter themselves into likeminded political communities provides us with an interesting opportunity to contrast Reddit users of differing ideologies with relative ease. We would like to study the qualities of discourse across different political ideologies on Reddit in this study. Of course, the attributes of online interaction are difficult to quantify in any meaningful, reasoned way. For this reason, we chose to use sentiment analysis, "a type of data mining that measures the inclination of peoples opinions through natural language processing (NLP), computational linguistics and text analysis" [8]. Our chosen sentiment analysis algorithm will score the positivity, negativity, and neutrality of each sampled comment, all of which will then be combined to create a single index measuring the net positivity of the comment (more on this later). Both sides of the aisle often accuse the other of being more interested in partisan negative attacks than a reasoned discussion of policy ideas; in this study, we therefore hope to gain some clarity on this issue and ascertain whether there is a statistically significant difference in the sentiments of liberal and conservative Reddit users when discussing politics, providing insight into the differences in rhetoric and communication on both sides of the political spectrum.

2 Study Design

Our project will be predicated upon an observational study of the sentiment of various Reddit users on political forums. There will be no blocking or control group due to the

[1] These include communities dedicated to movies, ideologies, news, and cats among many, many others.

purely comparative nature of the study.

2.1 Variables

Our explanatory variable is a categorical classification of one's political ideology as liberal or conservative[2]. To assign this classification, we use the partisan subreddits that the user is participating in as an accurate (albeit not infallible) heuristic to predict his/her ideology. Our response variable is the sentiment score of comments on political communities, measuring the positivity of a liberal or conservative user's comment.

2.2 Question and Hypotheses

> *Question:* Is there a difference in the mean sentiment (positivity) of liberal and conservative users when engaging in political discussion?

Because we start with no preconceived notion of which side of the ideological spectrum would tend to be more negative or positive and simply wish to measure *a* difference, we will conduct a two-sided t-test for a difference of means (mean sentiment scores in this case). As such, our null and alternative hypotheses can be stated as

$$H_0 : \mu_L - \mu_C = 0 \qquad H_a : \mu_L - \mu_C \neq 0 \tag{1}$$

where μ_L is the mean sentiment score of liberal users and u_C is the mean sentiment score of conservative users.

[2] "Liberal" and "conservative" are used as shorthand in this proposal for being on the left and right sides of the political spectrum, respectively.

3 Procedure

3.1 Sampling

Our population in this case is Reddit users on both sides of the ideological spectrum who participate in political discussion. Sampling on Reddit is a uniquely difficult problem because of the sheer number of communities dedicated to political discussion and activism, meaning that there are no two political subreddits that lend themselves to being perfect foils of each other. Another issue is the fact that Reddit as a whole skews liberal owing to its largely young userbase, meaning that ostensibly neutral subreddits must be treated as liberal havens[3] to accurately sample political discussion. For this reason, we ruled out a simple random sample, and chose instead to use a stratified design to attain a more complete depiction of both liberals and conservatives.

3.1.1 Stratified Random Sample

Ideological Classifications of Subreddits	
Liberal	Conservative
r/politics	r/Conservative
r/neoliberal	r/The_Donald
r/LateStageCapitalism	r/libertarian
r/esist	r/AskThe_Donald

Table 1: Subreddits from both sides of the ideological spectrum to be sampled from in this study. These communities represent different parts of the political community on each side and will be sampled from according to their size.

Table 1 shows the classifications we have determined for our samples. Though we have yet to determine the exact proportions, each community will be sampled from proportional to their size relative to the rest of the communities representing their ideology to ensure we

[3] Note that this is not purely arbitrary. r/politics, the largest discussion community with no stated ideology, has been consistently observed to have a definite leftward bias.

create a representative sample ($n = 500$ for each side, cumulatively). The four subreddits selected on each side are the largest political discussion (i.e. not simply memes or jokes) communities on Reddit with definite ideological bents.

3.2 Collection and Processing

To collect comments from each of these communities, we will collect a stream of all new comments with more than three words (to ensure the accuracy of our sentiment analysis) on all of the communities listed in Table 1 over the course of three consecutive days. This will be achieved with a query of Reddit's API using PRAW, the Python Reddit API Wrapper [9]. After this process is completed, collected comments will be filtered by the community they were posted in and sorted into their respective ideological assignments[4]. From there, each comment will be numbered and then a process of random selection (with a maximum number of comments from each community depending on its size to preserve the stratified sampling method) without replacement will take place until 500 comments have been selected from both liberal and conservative users.

Once the sample has been created, we will query the Parallel Dots Sentiment Analysis API [10] with each of the thousand comments to determine its positivity, neutrality, and negativity scores[5]. We will then subtract the negativity score from the positivity score to create a net sentiment score, which will range between -1 and +1 for the most negative and positive comments, respectively.

After all samples have been assigned a sentiment score, we will calculate the mean sentiment score for each ideology and proceed with our statistical inference test and confidence interval.

[4] We estimate that we will collect around 800,000 comments over the three day period if historical trends hold.

[5] These values are given as decimal values between 0 and 1.

4 Inference Test and Confidence Interval

We are conducting a two-sided two-sample t-test and interval for a difference of means. To be able to conduct this test, both samples must meet three criteria: randomness, independence, and normality. As discussed in Section 3.2, the samples will be randomly selected in a stratified process from the collected population of comments. The independence condition is met when $N \geq 10n$. In our test, we will have an N of approximately 800,000 comments with both samples being $n = 500$. As $800000 \geq 10(500)$, we conclude that the condition of independence is met for sampling without replacement. Finally, the normality condition can be met as long as $n \geq 30$ for both samples, which is indeed the case. We additionally expect the sentiment scores to be approximately normally distributed, but this is not necessary as the condition is already met.

We will perform the test and create the interval at 90% significance, or $\alpha = 0.10$ for the sake of keeping the power of the test relatively high.

4.1 Descriptive Statistics

When we report our results, we will provide the mean sentiment scores for both ideologies in a bar graph and provide all relevant breakdowns of the data in tables and graphs as needed. We will additionally report the sample standard deviation and the test statistic, as well as any other relevant summary statistics. At this time, it is difficult to state what kinds of graphs and tables will be provided in any further detail without first conducting the study and obtaining the data.

References

[1] *Political Polarization in the American Public — Pew Research Center*. 2019. URL: https://www.people-press.org/2014/06/12/political-polarization-in-the-american-public/ (visited on 04/12/2019).

[2] *Fewer now have mix of liberal, conservative views in U.S. — Pew Research Center*. 2017. URL: https://www.pewresearch.org/fact-tank/2017/10/23/in-polarized-era-fewer-americans-hold-a-mix-of-conservative-and-liberal-views/ (visited on 04/12/2019).

[3] Changjun Lee, Jieun Shin, and Ahreum Hong. "Does social media use really make people politically polarized? Direct and indirect effects of social media use on political polarization in South Korea". In: *Telematics and Informatics* 35.1 (Apr. 2018), pp. 245–254. ISSN: 0736-5853. DOI: 10.1016/J.TELE.2017.11.005. URL: https://www.sciencedirect.com/science/article/abs/pii/S0736585317305208.

[4] Christopher A Bail et al. "Exposure to opposing views on social media can increase political polarization." In: *Proceedings of the National Academy of Sciences of the United States of America* 115.37 (Sept. 2018), pp. 9216–9221. ISSN: 1091-6490. DOI: 10.1073/pnas.1804840115. URL: http://www.ncbi.nlm.nih.gov/pubmed/30154168%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6140520.

[5] Rishab Nithyanand, Brian Schaffner, and Phillipa Gill. "Online Political Discourse in the Trump Era". In: (Nov. 2017). arXiv: 1711.05303. URL: http://arxiv.org/abs/1711.05303.

[6] Justin Cheng et al. "Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions." In: *CSCW : proceedings of the Conference on Computer-Supported Cooperative Work. Conference on Computer-Supported Cooperative Work* 2017 (2017), pp. 1217–1230. DOI: 10.1145/2998181.2998213. URL: http://www.ncbi.nlm.nih.

gov/pubmed/29399664%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5791909.

[7] Seth Fiegerman. *Aliens in the valley: The complete and chaotic history of Reddit.* 2014. URL: https://mashable.com/2014/12/03/history-of-reddit/%7B%5C#%7DG1ZjPcbtFSq3 (visited on 04/12/2019).

[8] *What is Sentiment Analysis? - Definition from Techopedia.* 2019. URL: https://www.techopedia.com/definition/29695/sentiment-analysis (visited on 04/12/2019).

[9] Bryce Boe. *PRAW: The Python Reddit API Wrapper.* 2019. URL: https://praw.readthedocs.io/en/latest/ (visited on 04/12/2019).

[10] *SENTIMENT - ParallelDots Text API.* 2019. URL: http://apis.paralleldots.com/text%7B%5C_%7Ddocs/index.html%7B%5C#%7Dsentiment (visited on 04/12/2019).