Battleground State:
Characterizing and Understanding Political Bias in News Discourse on Reddit[*]

Mansi Ganatra
University of Southern California
Computer Science Department
Los Angeles, CA 90089
mganatra@usc.edu

Samar Haider
University of Southern California
Computer Science Department
Los Angeles, CA 90089
samarhai@usc.edu

Abstract

Increasing volumes of online content have resulted in social media websites and news aggregators becoming the primary means of information dissemination for a majority of Internet users. However, this has also lead to the creation recent political events have led to a rise in polarization in online communities and the formation of echo chambers that stifle political discourse across the ideological and party boundaries. There is a need to understand these phenomena in order to make efforts towards bridging this divide. In this project, we look at different types of biases that occur in discourse centered around news stories and how users respond to coverage from media on different ends of the political spectrum, as well as how these users interact with each other on particularly divisive topics. We use popular the news aggregation and rating website Reddit as our test bed, and collect submissions and comments from its highly active news and politics communities to conduct our analysis.

Introduction

While traditional brick-and-mortar newspaper readership has been seen to be on the decline, growing numbers of people have turned to the Internet for their news and updates. This has resulted in steadily increasing traffic to online platforms, where users can not only find the latest news but also respond to it and interact with others in real time.

Yet this increase in interaction between people with opposing views has not led to a decrease in the political polarization of communities, but has actually exacerbated the problem by providing users with tailored newsfeeds that silo people's opposing views from each other. The ramifications of this were seen most strikingly in the lead up to and the aftermath of the 2016 U.S. Presidential Elections, which saw a supposedly shock win of a candidate who did not appear to have garnered much support, at least from the Internet's perspective. This campaign showed that there is a need to analyze online political conversations to track the sentiment and polarity of different communities to better understand the momentum of the populace at large.

While most prior work (Flaxman, Goel, and Rao 2016) has focused on analyzing the political conversation on Twitter, relatively less attention has been paid to political conversations on Reddit, a social news aggregator and one of the most popular online discussion forums. Reddit ranks number 1 by number in the Alexa website rankings in the news category (Alexa 2019). Reddit is composed of a large number of communities called 'subreddits' where people react to and discuss various topics of interest. In (Mills 2018), an analysis of partisan subreddits centered around individual candidate for the U.S. Presidential Elections is conducted. However, these subreddits are, by definition, bound to be biased towards their respective candidate.

In this work, we aim to analyze trends and uncover biases in the online discourse in the wider news and politics subreddits, **r/news**, **r/worldnews**, and **r/politics**, which have around 45 million users in total. Figure 1 shows a breakdown of activity with respect to subscribers, posts, and comments for each of these subreddits over the course of the past few years. These communities are part of the default subreddits that new users are automatically subscribed to and are part of the list of subreddits that make up the Reddit homepage. As such they are frequented by people from across Reddit's user community and of all types of political leanings. We believe that an analysis of these communities would offer deeper insights into bias in online political discussions and help aid the understanding of news discourse on social platforms.

In the following sections we discuss related work, present our methodology and some preliminary analysis, and outline a plan for moving forward.

Related Work

A lot of effort has been put towards studying two fields - media bias and social media bias. Detailed analysis of effects of political fragmentation caused due to bias in traditional news media outlets has been discussed in (Morris 2007). The paper provides evidence of how news channels benefit from such a fragmented viewership. On the other spectrum of social media, the growing field of personalization has made it

[*]Note that this title differs from that specified in the proposal due to a late pivot in the project topic due to unforeseen issues in data acquisition for the work initially planned in the project proposal.
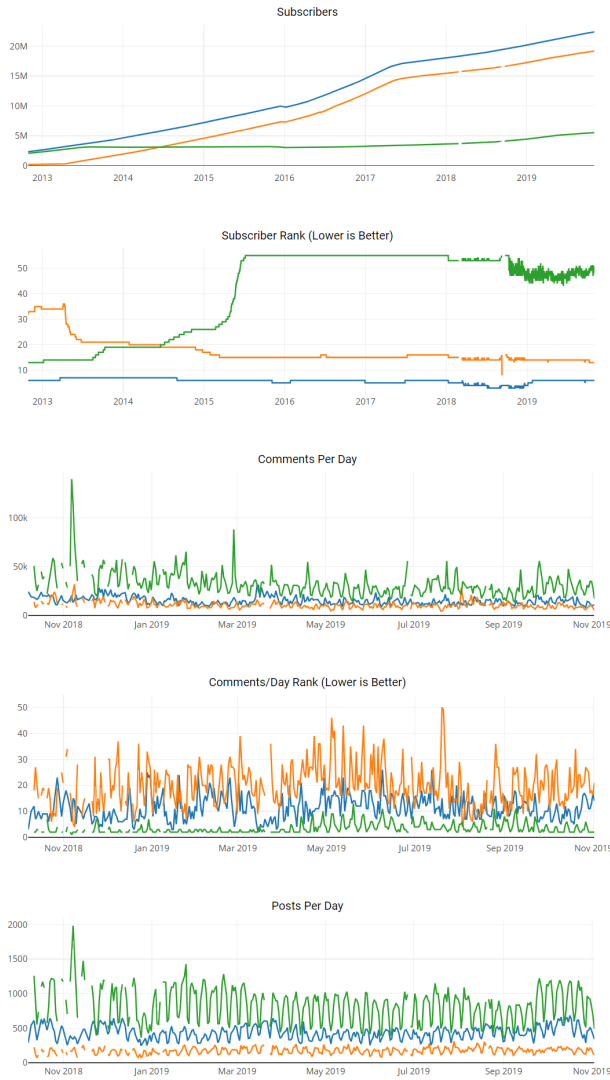
Figure 1: Activity statistics for the three subreddits r/worldnews (blue), r/news (orange), and r/politics (green) (Subreddit Stats 2019). All three subreddits have seen a steady growth in subscribers. However, it appears that r/politics sees more posts and comments per day, yet has a lower subscriber rank than the two larger communities.

r/worldnews	r/news	r/politics
theguardian.com	cnn.com	thehill.com
independent.co.uk	apnews.com	washingtonpost.com
bbc.com	nbcnews.com	newsweek.com
reuters.com	theguardian.com	independent.co.uk
bbc.co.uk	reuters.com	cnn.com
thehill.com	cbsnews.com	businessinsider.com
newsweek.com	foxnews.com	commondreams.org
cbc.ca	abcnews.go.com	thedailybeast.com
businessinsider.com	nytimes.com	cnbc.com
cnbc.com	bbc.com	nytimes.com
telegraph.co.uk	cnbc.com	thinkprogress.org
france.com	independent.co.uk	self.politics
dw.com	bbc.co.uk	salon.com
commondreams.org	thedailybeast.com	lawandcrime.com
news.sky.com	cbc.ca	vox.com

Table 1: Top 15 news websites posted in each of the three subreddits (Subreddit Stats 2019). We observe some clear differences in source popularity between the three communities, with r/news sand r/politics skewed more strongly towards North American outlets in comparison to r/worldnews' more balanced list.

Reddit: votes, likes, users' browsing history, typical voting behavior, etc.

As more and more social media institutions adopt content personalization, there are more questions arising on the strategies and methodologies employed for it. (Bozdag 2013) discusses in extensive details about the role of filtering in the personalization algorithms and how this filtering is influenced by humans at each step. Research has proved that traditional media has its own biases in data gathering and fact checking and how some of these biases are conscious choices to drive the audience numbers. Research continues to provide extensive evidence of algorithmic biases and its impact on humans in real life. Also, increasing number of social media companies now act as curators for content- all forms of content. So the question now arises of how all of the above three factors influence on the way we consume news in the 21st century?

Methodology

Data

The first stage in our project is the collection of data. For the purpose of this project, we focus on Reddit submissions and comments centered around the 2016 U.S. Presidential Elections. We narrow our search down to the one month before, the month of, and the month after the elections. For collecting this data, we have been considering multiple strategies. The first is through a Reddit API that allows us to query the website for posts and comments that match a certain criteria. However, this API is rate-limited and so we look towards an online repository of all historic Reddit posts, pushshift.io, for a complete dump of the required data. This is a large amount of data that we are currently working with and attempting to filter into the exact subreddits and time-frame that is our focus. Figure 2 shows an example JSON response

imperative for social media websites to curate content specific to user. This personalization serves two main purposes: i.) Content curated for specific users give them the ease and convenience to easily access the information they want to consume, and ii.) It helps the social media company to effectively sort through the influx of information and show only certain amount of it to user, saving big on resources. The era of social media personalization started with its large scale adoption by Netflix. (Amatriain 2013) discusses in detail the models used by Netflix for personalization, architectures for systems streaming real-time personalized content and handling data flow in such systems. On the same lines, (Glenski and Weninger 2017) focuses on predicting user behavior on

received by querying the API for posts from the r/news subreddit.

```
        "posts": [
            {
                "id": "do73ja",
                "author": "KinkyQuesadilla",
                "created_utc": 1572259323.0,
                "distinguished": null,
                "edited": false,
                "locked": false,
                "name": "t3_do73ja",
                "num_comments": 1399,
                "permalink": "/r/news/comments/do73ja/missouris_only_abortion_clinic_to_challenge_state/",
                "score": 16886,
                "selftext": "",
                "upvote_ratio": 0.92,
                "url": "https://www.reuters.com/article/us-usa-abortion-missouri/missouris-only-abortion-clinic-to-challenge-state-shutdown-order-idUSKBN1X70WT",
                "comments": [
                    {
                        "id": "f5le14m",
                        "author": "DuskTrillDawn",
                        "body": "I used to live like 2 streets down from this clinic. There were ALWAYS protesters out there screaming at people and they made everyone feel uncomfortable who was going in or just walking near there.",
                        "created_utc": 1572277322.0,
                        "distinguished": null,
                        "edited": false,
                        "is_submitter": false,
                        "link_id": "t3_do73ja",
                        "parent_id": "t3_do73ja",
                        "permalink": "/r/news/comments/do73ja/missouris_only_abortion_clinic_to_challenge_state/f5le14m/",
                        "score": 1623,
                        "subreddit_id": "t5_2qh3l"
                    },
```

Figure 2: An example of the JSON response received from querying the Reddit API for posts from the r/news subreddit.

Media Bias Analysis

To cluster news outlets by their political stance, we use the the media bias ratings provided by (AllSides 2019), which divide the most popular news websites by how left- or right-leaning they are. A preliminary analysis of the top 15 most popular URLs posted in each of the three subreddits we focus on is given in Figure 1.

Figure 3: Which way do you lean? The independent media rating body AllSides provides a detailed breakdown of popular media outlets on a 5-point scale based on a multi-partisan scientific analysis of their online content (AllSides 2019).

Future Work

In the coming weeks we intend to perform an extensive analysis of the data collected, including, but not limited to, the following steps:

1. **URL + Score Analysis**
   - Analyse most frequent websites in posts
   - Cluster websites by political leaning using (AllSides 2019)
   - Find average scores by website and political leaning
   - Analyse the cross-posting of websites across the political spectrum

2. **User Response Analysis**
   - Analyse which users respond to which posts based on political leaning
   - Cluster users based on the frequency of their response to posts across the political spectrum
   - Analyse controversiality measure of posts for each political leaning

3. **Text Mining and Sentiment Analysis**
   - Extract features from comments on posts
   - Analyse polarity of users for/against political sides
   - Cluster users based on the sentiment of their response to political sides
   - Analyze the sentiment of users cross-posting on different political stances.

References

Alexa. 2019. Top sites by category.

AllSides. 2019. Media bias chart.

Amatriain, X. 2013. Big & personal: data and models behind netflix recommendations. In *Proceedings of the 2nd international workshop on big data, streams and heterogeneous source Mining: Algorithms, systems, programming models and applications*, 1–6. ACM.

Bozdag, E. 2013. Bias in algorithmic filtering and personalization. *Ethics and information technology* 15(3):209–227.

Flaxman, S.; Goel, S.; and Rao, J. M. 2016. Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly* 80(S1):298–320.

Glenski, M., and Weninger, T. 2017. Predicting user-interactions on reddit. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, 609–612. ACM.

Mills, R. A. 2018. Pop-up political advocacy communities on reddit. com: Sandersforpresident and the donald. *AI & SOCIETY* 33(1):39–54.

Morris, J. S. 2007. Slanted objectivity? perceived media bias, cable news exposure, and political attitudes. *Social Science Quarterly* 88(3):707–728.

Subreddit Stats. 2019. Subreddit stats.