Reddit Social Analysis With Topic Modeling

Carissa Bleker[1,2], Ashely Cliff[1,3], Sean Whalen[2], and Saeed Beztchi[2]

I. Introduction

Reddit (`reddit.com`) is a social website where users congregate online to share, vote on, and discuss content. Currently Reddit is the 3rd most popular social website in the US, and the 6th most popular worldwide [1]. Reddit provides numerous forums or communities, termed subreddits, that cover many unique and overlapping interests. These can range from news, sports, politics, hobbies, entertainment, etc. Each of these subreddits may have tens to millions of users submitting topic specific content, creating discussion threads on these submissions, and either "up-voting" or "down-voting" specific submissions or comments.

As a popular website, Reddit has a large impact on how information is shared, and how perspectives or opinions on topics are formed and adopted by users [2]. Analysis of the submissions, comments and voting results should help expose these patterns. With the data and methods discussed below, we aim to provide insight into how topics included in subreddits evolve over time, how Reddit users perceive certain topics, and how users associate with certain topics.

II. Data

The hierarchy of Reddit is shown in Fig 1. Reddit contains numerous subreddits focusing on specific areas of interest. Users can submit content to these subreddits, and comment on each others submissions.

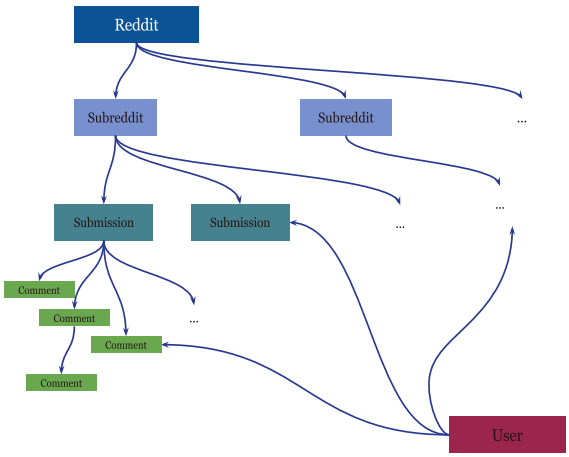

Fig. 1. Reddit hierarchy.

[1]The Bredesen Center for Interdisciplinary Research and Graduate Education, University of Tennessee, Knoxville
[2]The Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville
[3]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN

Reddit has a dedicated API for interacting with the website and downloading data. The Python Reddit API Wrapper (PRAW) [4] provides a python interface to the API, and PRAW was used to gather a majority of the data used.

Google BigQuery was also used for some of the data collection, in particular associating Reddit submissions with the day they were created. This was done by selecting the `link_id` and `created_utc` fields of Reddit dumps made available by Google. We used the BigRQuery R package to query the tables from Rstudio and saved the resulting tables locally.

The collected data was gathered from two distinct subreddits. A years worth of data (October 2016 through September 2017) for /r/worldnews, and six months worth of data (January 2017 through June 2017) for /r/nba was collected. Data gathered included submission IDs, submission authors, submission comments, submission dates, and comment authors. We saved the comment forest for each submission as a single document.

III. Methods

Topic Modeling

We used Non-Negative Matrix Factorization (NMF) to identify topics within the documents. NMF is a method of dimension reduction and latent factor extraction that is commonly used for topic modeling in text. For an input matrix $X_{n\times m}$, NMF finds $W_{n\times t}$ and $H_{t\times m}$ that minimize

$$||X - WH||_F,$$

where $||\cdot||_F$ is the Frobenius norm, and therefore $X \approx WH$. In our setting, X is the input matrix with n documents and the counts of m words in each of these documents.

Each row of H corresponds to a *topic*: a collection of words that contributes to the distribution of words in the document. Each entry $w_{i,j}$ of W gives the contribution of topic i in H to document j in X. We used $m = 1000$ and $t = 10$. We calculated the topics using the documents per month and over the whole year.

Aggregated Topics

Similar topics appeared in different months in the year. Certain topics also appeared and/or disappeared as the year progressed. In order to follow these topic developments, we needed to match similar topics found in different months. To this end we calculated the overlap in the top 20 words for each topic, across months. Those with 40% or more overlap were deemed to be the same topic.

Controversial Topics

Reddit uses tags to describe user interactions with a submission. These tags include *Hot*, *New*, *Rising*, *Controversial*, and *Top*. By selecting a tag, the submissions within a subreddit will be ordered accordingly. We were particularly interested in submissions that ranked high in Controversial. Ostensibly, these are submissions that received an equal number of upvotes and downvotes, and therefore generated divided feelings from Reddit users.

To determine if certain topics are more likely to be controversial, we compared the topic contributions in the highly ranked submissions to those in the rest of the submissions.

Suppose C is the set of controversial submissions, then for each topic i we calculated:

$$r_i = \frac{\sum_{j \in C} w_{i,j}}{\sum_{k=1}^{K} \sum_{j=1}^{|C|} w_{i,j}^{*k} / K},$$

where w^{*k} is based on a k^{th} random selection of $|C|$ submissions.

Users

We hypothesized that users will tend to participate in communities based on their similarities, or gravitate to specific topics within a subreddit. Using the results from topic modeling, submissions can be mapped to specific topics, and through the submission and comments, users can be associated with these same topics.

We constructed bipartite networks of submissions, topics, and authors from the above analysis, and explore a number of patterns we discovered in these graphs.

IV. Results

A. World News

We performed topic modeling on the the most popular news subreddit (/r/worldnews). We were able to assign each topic a descriptive name. Table I shows the topics defined for the year October 2016 to September 2017. Many of these topics also appeared throughout the month topics. In the year and in each month we noticed a single topic that appeared to have no meaning. We assigned this topic the name *Reddit*, as it appeared consistently in the analyses.

TABLE I
World News Topic Words

Topic	Words
Reddit	money work government life pay drug
North Korea	nk korea north nuclear kim korean war
Russia	russia russian putin ukraine nato russians
Israeli-Palestinian conflict	israel palestinians israeli jews palestinian
US politics	trump president obama election hillary
Climate Change	climate energy solar coal change ice
EU/Brexit	eu brexit uk vote referendum germany
Asia	china chinese india japan trade countries
Syria Conflict	isis iran turkey saudi assad syria war iraq
Religion	muslims muslim islam religion women

Topic aggregation resulted in 23 topics, some of which occurred in multiple months, and others that only occur in one month. Fig 2 shows the aggregated topics over the year.

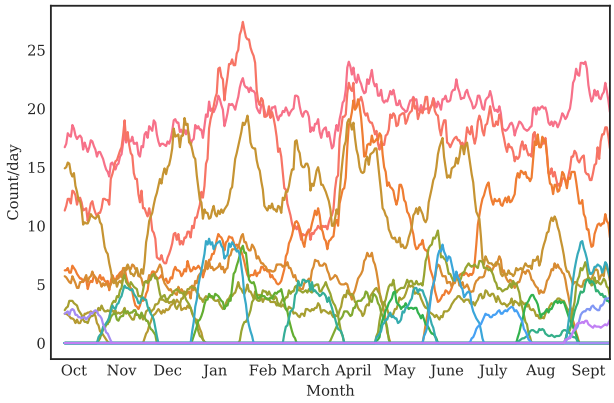

Fig. 2. Aggregated topics per day for the year (smoothed by moving average).

We were able to quantify a number of the topic appearances over time with news events of the same month. In particular, a few example aggregate topics are shown in Fig 3, annotated with names derived from the most common topic words. We also noticed certain topics slightly changing over

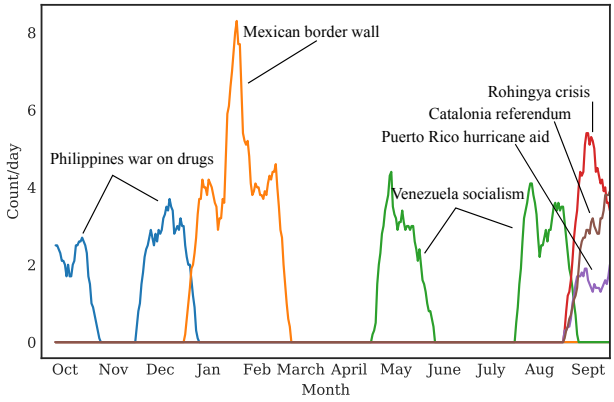

Fig. 3. Annotated topics (smoothed by moving average).

time with respect to news events. An example of this is the appearance of the word *hurricane* in the *Climate Change* topic in the month of Hurricanes Irma and Maria, and Jose (September 2017). This pattern was also prevalent in the *US politics* topic.

B. NBA

Similar analysis was performed on the NBA subreddit (/r/nba). This subreddit was deliberately chosen to contrast the seriousness of /r/worldnews, and because the NBA has a clearly defined schedule with which we can link real world events to topic appearances. Table II shows the first four words of the ten topics generated for the NBA subreddit during the second half of the 2016-2017 season (January-June 2017).

TABLE II

NBA TOPIC WORDS

Topic	Words
0 *n/a*	com https fuck guy http
1 *Cavs*	lebron cavs kyrie finals cleveland
2 *Mvp Race*	harden westbrook russ mvp triple
3 *Trades*	trade pick draft picks lakers
4 *Goats*	kobe shaq jordan player lebron
5 *Playoffs*	celtics cavs warriors series playoffs
6 *Warriors*	kd warriors curry steph durant
7 *Spurs*	kawhi spurs pop lma manu
8 *Refs Suck*	foul refs wall ball shot
9 *Knicks*	melo knicks phil kp dolan

On examination, topics *n/a* and *Refs Suck* appear to be meaningless; we could not find a connection to tie the words together. Topic *Refs Suck* could be about NBA referees but 'wall', a player for the Washington Wizards, is oddly included.

Topics *Cavs*, *Warriors*, *Spurs* and *Knicks* are clearly about NBA teams: the key words comprise personnel present on each respective roster. Topic *Mvp Race* includes the words *harden* and *westbrook*, the MVP and MVP runner-up for the season. Topic *Goats* is about hall-of-fame caliber past players. Topic *Playoffs* includes *cavs*, *warriors*, and *celtics* - all playoff teams - and *series*, so it is likely a playoff series topic. Fig 4 shows a heatmap of the above topics over time.

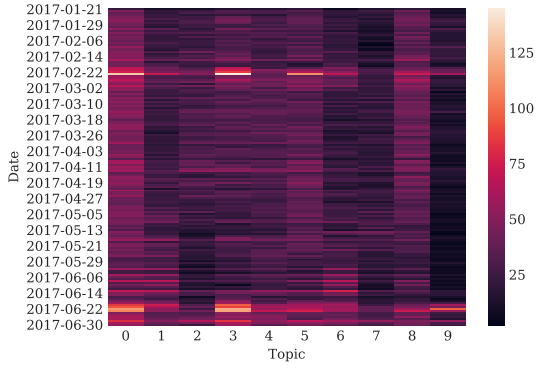

Fig. 4. NBA topic occurrence over time.

After examining the heatmap we became interested in what occurred during the hotspots for topic *Trades* (3), which included the key words *trade*, *pick*, *draft*, and *picks*. We plotted it against time as a line graph in Fig 5. We found that the peaks at 2/23 and 6/22 line up exactly with the NBA 2016-2017 trade deadline and draft, which demonstrated to us that our analysis was meaningfully tracking discussions in response to real-world events.

C. Controversial Topics

Fig 6 displays the results of the controversial analysis within /r/worldnews for the year October 2016 to September 2017. Topics with a ratio below 1 are less likely to occur in a controversial submission, and topics above one are more likely to appear in submissions labeled as controversial. The topics *North Korea*, *EU/Brexit* and *Asia* are less likely to

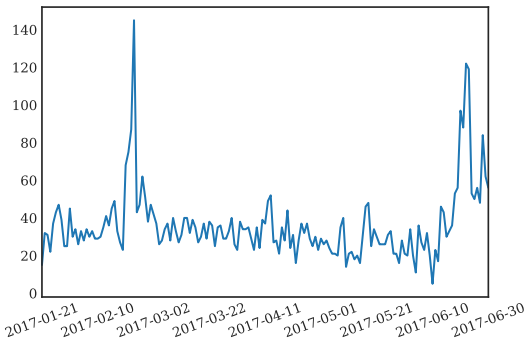

Fig. 5. NBA *Trades* topic occurrence.

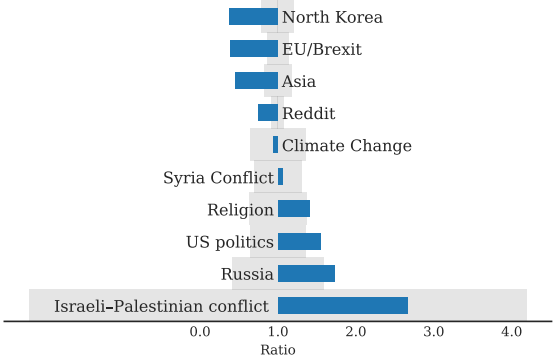

Fig. 6. Ratio of topic contributions in controversial submissions to background submissions. The shaded area is the standard deviation of the background samples.

be controversial, while *Religion*, *US politics Russia* and the *Israeli-Palestinian conflict* are more likely to be controversial.

Given that the majority of Reddit users are US based, these results seem unsurprising.

D. Users

Figures 7,8, and 9 show three different views from the user results. The user analysis was done using the topics results and Reddit data for October of 2016.

Figure 7 shows the users that are associated with topic seven from October of 2016, which was the Israeli-Palestinian conflict. Each edge represents a different submission or comment from the user (in blue) to the topic (in red). Notice that a few users have a very large number of submissions/comments for a single month. Only users that had at least one submission were included in the calculations, as many users leave only comments - which is deemed less of a commitment, and removal of these reduced the amount of data to a reasonable amount and also removed less relevant users.

Figure 8 shows users connected to associated topics via only submissions, not comments. Notice that some authors tended to cluster around certain topics. However, many users did not cluster and had submissions associated with different

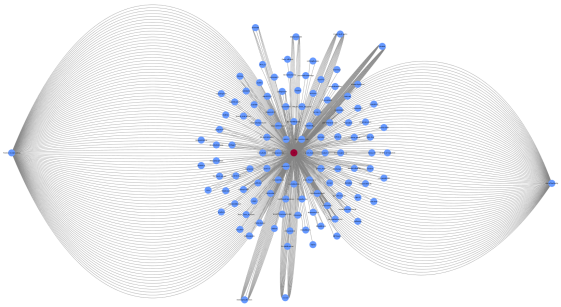

Fig. 7. Submission authors to a topic.

topics, which was not expected.

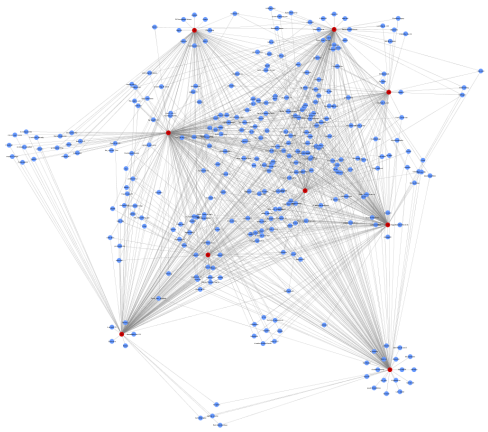

Fig. 8. Submission authors to topics.

Figure 9 shows users associated with topics via comments and submissions, where the edge thickness corresponds to the number of comments and submissions created by that user associated with that topic. Notice that some users have thick edges to many topics, indicating that they simply comment and submit about anything and everything, however some have much thicker edges associated with some topics and much thinner edges with all others. This shows that while some users may not associate with any one topic, many others may associate more so with one than others, but not purely.

User analysis and networks created from the topics and users can be used find authors that have an agenda - either the push or suppress any specific topic. Users that have a large number of submissions falling solely within one topic, but no comments or submissions in others may be a bot, and could be flagged - as spamming is against Reddit's terms of service.

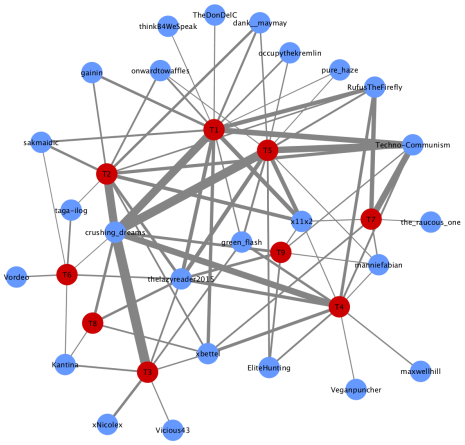

Fig. 9. Submission and comment authors to topics.

V. Problems

While using an API and an API wrapper makes programming and data management more user friendly, it did include some problems. Such factors included obtaining a shorter history of a subreddit's comments, and limited history available on submission tags. This means for further work we would need to collect data at perhaps a daily scale. Another issue with data collection was deleted information (either a comment or user), which in all cases we simply ignored.

Due to the sheer number of comments that a submission may have, the number of comments used for analysis was limited to the unexpanded list of comments the Reddit API provides. The same constraint was used when looking at comment authors for the user analysis.

We also did not take into account the parameter values we chose for analysis. This includes the minimum number of comments per submission, the number of words used per document, the number of topics, and thresholds we used in determining aggregated topics.

VI. Conclusion and Future Work

Through the analysis process we applied, we were able to computationally track the rise and fall of meaningful topics on Reddit, within specific subreddits. We learned that contrary to our initial hypothesis, users do not simply cluster around topics, but rather may connect highly with many or with a few and less so with others. We were also able to indicate which topics are considered controversial, and thus determine which topics divide people, at least within the Reddit community.

Future work may involve applying the user analysis to find users or bots pushing certain agendas. Work may also be done to identify sides of issues once topics have been determined.

These analyses could be applied to other subreddits, as well as to many other social media platforms to find topics and connections within many different subpopulations. This type of information may prove useful when tracking news, news consumption and development of opinions within and across online communities.

REFERENCES

[1] SimilarWeb.com. 2017, August. Top Social Network Websites in United States. [online] Available at: https://www.similarweb.com/top-websites/category/internet-and-telecom/social-network [Accessed 27 Sep. 2017].

[2] Weninger, T., Zhu, X.A. and Han, J., 2013, August. An exploration of discussion threads in social news sites: A case study of the reddit community. In Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (pp. 579-583). ACM. Vancouver

[3] Mei, Q. and Zhai, C., 2005, August. Discovering evolutionary topic patterns from text: an exploration of temporal text mining. In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining (pp. 198-207). ACM.

[4] Boe B. PRAW: The Python Reddit API Wrapper. 2012-, https://github.com/praw-dev/praw/ [Online; accessed 2017-09-29].