Tech, Media and Democracy

Jingxuan Sun, js3422
Yue Wang, yw986

Conflicts between Subreddits and how Users Behave towards them?

March 23, 2019

1. Abstract

Through this project, we analyze the conflicts between subreddits. This is related to online community and “harmful” content as we discussed in class. First we use the dataset from Stanford to detect inter-subreddit conflict for a series of subreddits [1]. Conflict is defined by identifying the source and target subreddit. If a post from the source subreddit contains a link to the target subreddit and the content is malicious, the post is considered as an initiator of conflict.

After conflict identification, we further focus on the initiator post, and observe the characteristics (including activeness, karma, engagement) and behavior (majorly language usage) of the author and the commentators. In addition, we conduct sentiment analysis with both Perspective API and NLTK toolset on the post/submission author and commentators to have a coarse comparison between the two libraries.

2. Goal of Analysis

User-defined and user-organized communities are an essential component on Reddit, where people could express their free speech and share information. However, there also lies the potential that some users intend to cause conflict between users, communities or other targets. Even worse, they might intend to spread misinformation through conveying false or negative

contextual meaning based on a true content. (As we discussed in Reading “False News, It’s complicated [2]).

Therefore, on Reddit platform, we choose the conflict form as inter-subreddit conflict as our research target. Here the inter-subreddit interaction is defined as one submission including hyperlink towards another subreddit. And the contextual information is acquired by extracting the headline and body text from original submission. Users behavior is also taken into consideration, as we extract both submissions’ authors and comments’ authors information. Detailed method is explained in SECTION 3.

1. **Analyze the sentiment information under negative submission:**

   As we discussed in class, users behave diversely between true news and false news, as the false news would stir higher attention and engage more discussion. So would there also be a difference between the reaction to positive and negative submissions in subreddit?

   One thing to clarify is that, here we define submission as the action of posting under subreddit which includes a hyperlink to another subreddit. And the criteria of defining positive and negative sentiment is based on the contextual language being used inside the body/title of a post and is provided in **Subreddit Hyperlink Network**[3].

   After extracting all the negative submissions, it is important to understand how would users react to negative post, would they be influenced and what kind of sentiment would be conducted in the comments? Here we intend to use PRAW API[5] to fetch more solid data and use two different sentiment analysis libraries, namely Vader in NLTK[6] and Google Perspective API [4] to verify.

2. **Analyze negative comments’ author behavior:**

   Are users involved in negative post discussion more active than others? If one express toxic comment under one negative post, would there has higher probability that he is more offensive than others?

Full code used can be accessed here [https://github.com/shakingkelly/TMD-reddit-conflict-analysis/blob/master/TMD.ipynb].

3. Data Collection

Data is complicated.

The primary dataset we used is **Subreddit Hyperlink Network**[3]**:** the subreddit-to-subreddit hyperlink network is extracted from the posts that incorporate hyperlinks from another subreddit in the original post content. We say a hyperlink originates from a post in the “source community” and links to a post in the “target community”. Each hyperlink is annotated with five attributes: the timestamp, source subreddit, target subreddit, the sentiment of the source community post towards the target community post (1 being neutral or positive while -1 being negative), and the text property vector of the source post. This dataset contains all the submissions under all subreddits from `2013-12-31 to 2017-04-30`, which contains 858,490 submission information.

571927 interactions were classified with Amazon Mechanical Turk, according to the study, and in general, over 89% of interactions were classified as positive or neutral, which left 6271 negative interactions. Our analysis is majorly carried out on these initiator posts.

As we need to analyze user behavior and portfolio, we use **PRAW** [5], an acronym for "Python Reddit API Wrapper" to crawl submission’s author and comments information below certain submission, and author of comments.

The data attributes we used are shown in the following table, where /id served as primary key for submission, and /commentsForest was used to trace down the comments and authors were found by /author.

Submission

Attribute	Description
`author`	Provides an instance of `Redditor`.
`clicked`	Whether or not the submission has been clicked by the client.
`comments`	Provides an instance of `CommentForest`.
`created_utc`	Time the submission was created, represented in Unix Time.
`distinguished`	Whether or not the submission is distinguished.
`title`	The title of the submission.
`upvote_ratio`	The percentage of upvotes from all votes on the submission.
`url`	The URL the submission links to, or the permalink if a selfpost.

Redditor

Attribute	Description
`comment_karma`	The comment karma for the Redditor.
`comments`	Provide an instance of `SubListing` for comment access.
`created_utc`	Time the account was created, represented in Unix Time.
`id`	The ID of the Redditor.
`is_mod`	Whether or not the Redditor mods any subreddits.

4. Analysis Approach & Result

1. **Sentiment analysis of comments under negative submission/post:**

After extraction of negative post IDs, we queried Reddit for post author, comments, upvotes. We tried to do the same for positive posts, but unfortunately all post IDs from the hyperlink dataset were invalid and we could not find the original pages for those posts.

We then shifted our focus to negative posts. We scraped all comments under each negative post and used NLTK for sentiment classification. Sample sentiment result from nltk.sentiment.vader:

	flirtation	identity_attack	idx	insult	profanity	sexually_explicit	text	threat	toxicity
0	0.519920	0.211680	0	0.264306	0.316808	0.412435	At first I typed out a comment about how butth...	0.280437	0.252440
1	0.403854	0.140421	1	0.236061	0.397672	0.799199	Where was the rape joke?	0.373892	0.501629
2	0.263957	0.046450	2	0.033025	0.044540	0.088662	SnapShots: [1] (http://archive.is/2twN9), [2] (h...	0.144010	0.054881
3	0.497973	0.134218	3	0.248090	0.438623	0.838079	No rape joke, not sure if disappointing or hap...	0.408097	0.506458
4	0.099253	0.057346	18959	0.042744	0.020241	0.033699	\n\nRemoved: Only post links if you are not th...	0.151483	0.064819

As shown in *Figure 2-1*, there is a balanced proportion of positive and negative comments under negative posts both at around 0.22, while the majority of comments are still neutral. Therefore, negative sentiment of original post title doesn't affect the sentiment distribution of the comments.

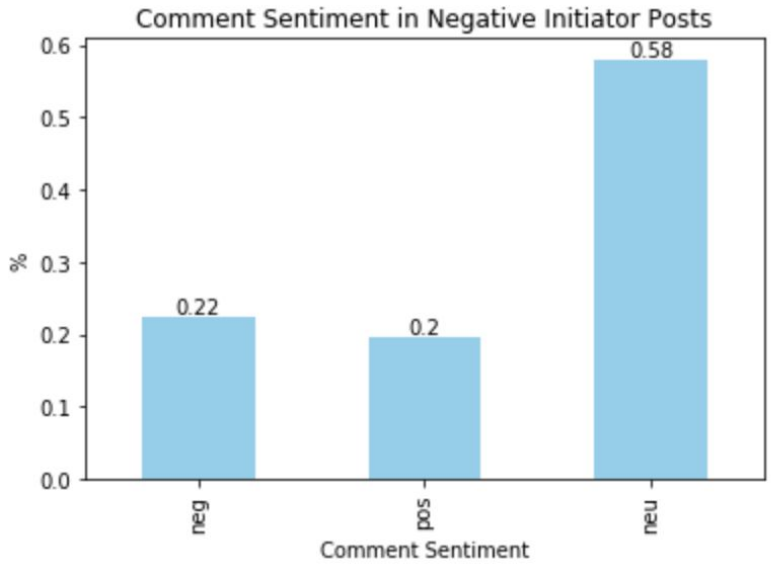


Figure 2-1:Comment Sentiment Analysis under Negative Post

Further, we used number of upvotes for each comment as a proxy of popularity/controversy of that comment. Especially we looked at the comments that received higher attention, referring to those received more than 100 upvotes or downvotes (-100 upvotes means 100 downvotes). We can observe that for positive and negative comments, they have a similar distribution of popularity. But the popularity of negative comments has a tighter distribution, which in turn means that in the positive distribution there are more outliers.

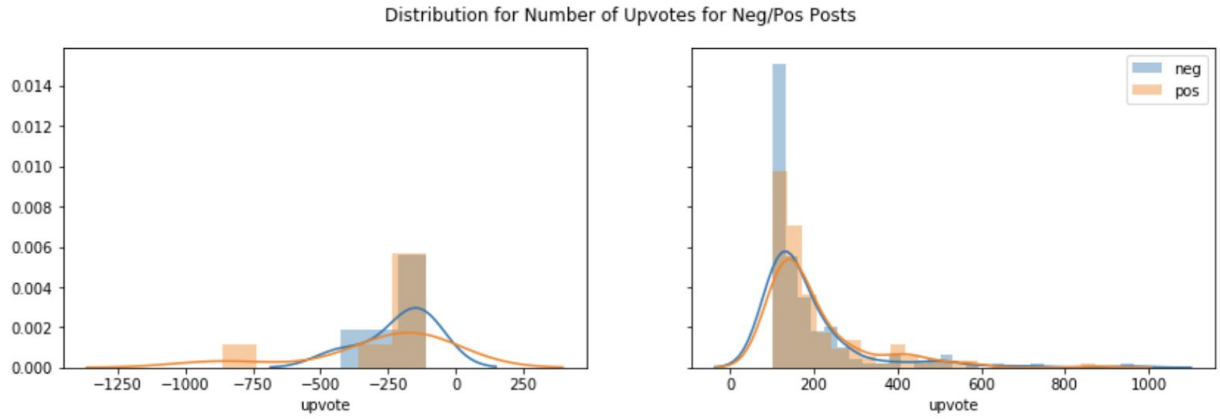


Figure 2-2: Upvote & Downvotes Distribution of Positive & Negative Comments

We have two hypotheses to account for the phenomenon. First, there might be accuracy limitations of the NLTK sentiment analyzer which would result in misclassification of comment sentiments and in turn the shift of positive/negative popularity distribution. For example, a neutral comment "Where was the rape joke?" (submission_id = '1u58yq') was classified as negative simple due to the appearance of "rape". Second, if the classification reflects the true sentiment of the comments, it might be that positive comments give narratives of controversial opinions, while there is a higher proportion of negative comments just expressing negativity through swear but not actually offering useful information to the discussion.

In addition, we complemented it with Perspective API for more fine-grained comment sentiment analysis, which contains six sub-categories[4]:

- DENTITY_ATTACK: negative or hateful comments targeting someone because of their identity.
- INSULT: insulting, inflammatory, or negative comment towards a person or a group of people.
- PROFANITY: swear words, curse words, or other obscene or profane language.
- THREAT: describes an intention to inflict pain, injury, or violence against an individual or group.
- SEXUALLY_EXPLICIT: contains references to sexual acts, body parts, or other lewd content.
- FLIRTATION: pickup lines, complimenting appearance, subtle sexual innuendos, etc.

Sample result is shown as following:

	author	comment	compound	is_submitter	neg	neu	pos	submission	submission_autho
0	ValedictorianBaller	At first I typed out a comment about how butth...	0.7278	False	0.000	0.810	0.190	1u58yq	sw33n3y
1	odintal	Where was the rape joke?	-0.5423	False	0.475	0.303	0.222	1u58yq	sw33n3y
2	ttumblrbots	SnapShots: [1] (http://archive.is/2twN9), [2](h...	0.0000	False	0.000	1.000	0.000	1u58yq	sw33n3y
3	CantaloupeCamper	No rape joke, not sure if disappointing or hap...	-0.6167	False	0.501	0.226	0.273	1u58yq	sw33n3y
5	NaN	Much Big Ten privilege. Many upset over those ...	0.8126	False	0.133	0.357	0.510	1u58yq	sw33n3y

The following graph shows the distribution of toxicity subtypes for all comments under negative posts. In agreement with *Figure 2-1*, no more than 20% comments were classified to be toxic in each subtype. The most common types of toxicity are profanity and insult, while flirtation, threat and identity attack are less observed.

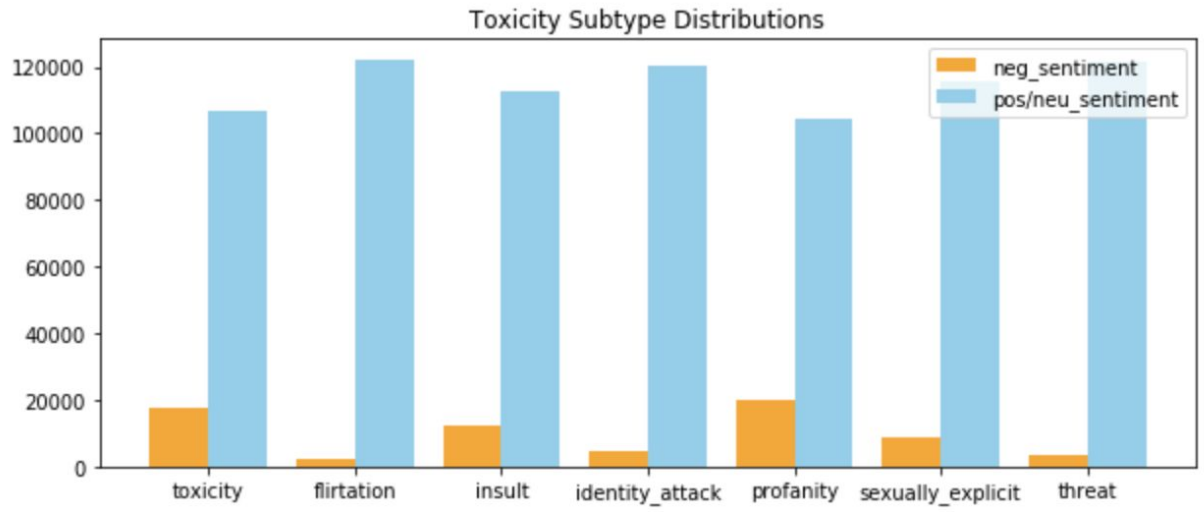


Figure 2-3: Distribution of toxicity subtypes among comments

Besides, a simple comparison between the two tools with the example "Where was the rape joke?" (submission_id = '1u58yq') shows that the performance of Perspective API resembles human judgement more.

2. **Negative comments' author behavior:**

We took a further step to analyze user behavior for the negative comment authors. To simplify the task, we scraped newest 100 comments for top 200 negative comment authors. We selected the top authors by total number of negative comments appeared in all the negative posts we collected. The top two author, autowikibot and totes_meta_bot, had 21 and 20 comments in the dataset, while the 200th author had 3, as shown in *Figure 2-1*. The attributes we used are whether the commentator is ever a **moderator** for any subreddit, **user age** (inferred from register timestamp and newest poactivity timestamp) and total **comment karma** as a proxy to activeness.

	count
autowikibot	21
totes_meta_bot	20
shitpostwhisperer	9
cordis_melum	8
deathpigeonx	8
EvanHarper	7
blarghable	7
5th_Law_of_Robotics	7
Slutlord-Fascist	7
SweetNyan	6

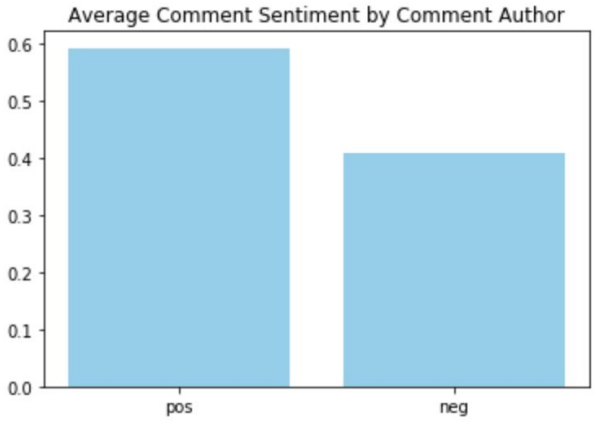


Figure 2-1: Top-negative comment author ***Figure 2-2:Average Comment Sentiment Analysis***

The most interesting result of this analysis is shown in the following graph. Nearly 70% of the top 200 commentators with negative sentiment have been moderators for certain subreddit. We haven't found statistics of moderator proportion for the whole platform, but we consider 70% is a high number.

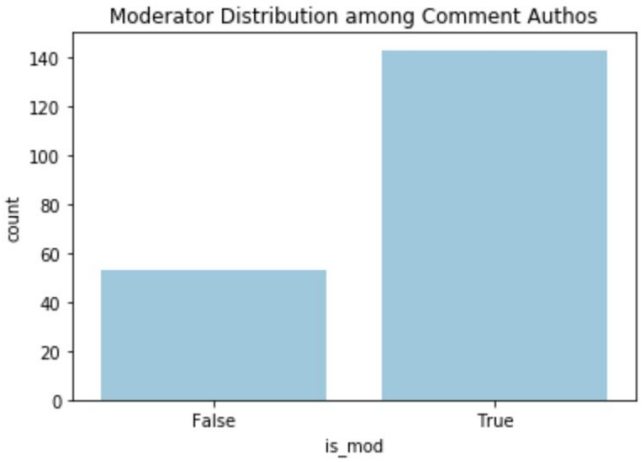
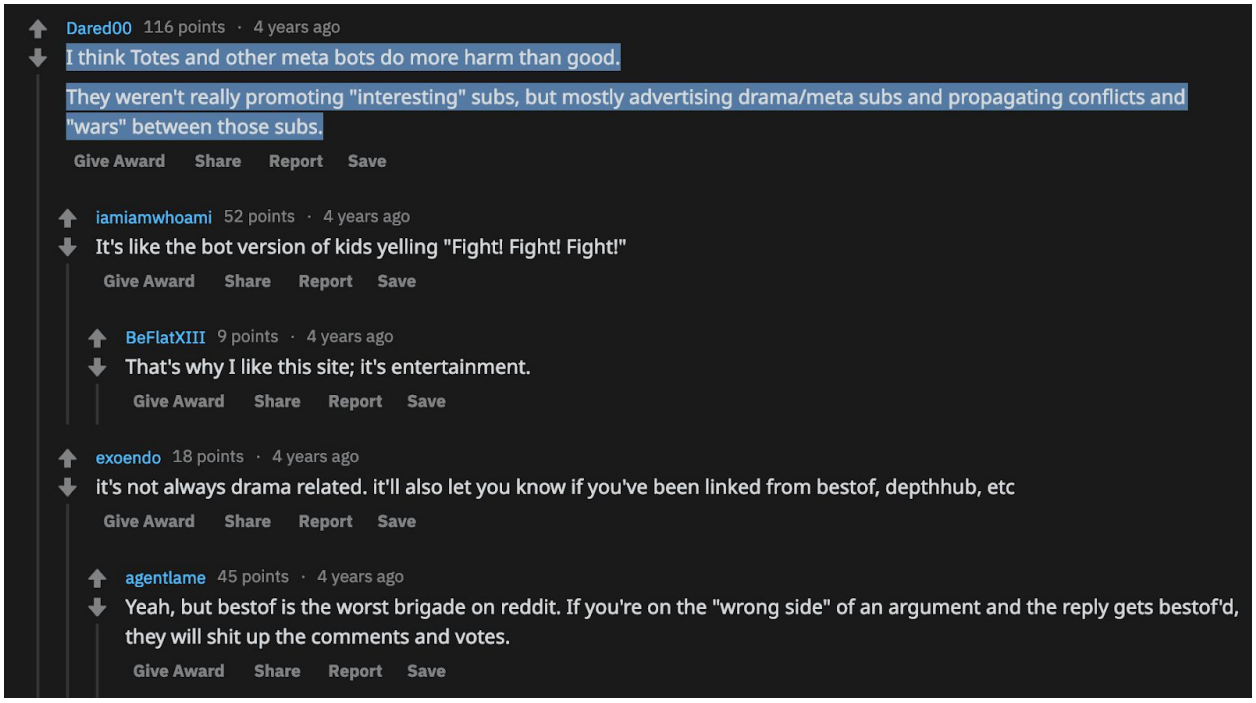


Figure 2-3: Determination of mode and non-mod among comment author

Accounting for it, I think it's partially due to the fact that there are multiple bot accounts in top 200 commentators as mentioned in the last paragraph, and bots in general have a high probability of being a moderator. However, it is hard to automatically identify bot accounts because there are both automatic bots and human-operated bots, and there is not a pattern of language use for the latter type. We might be able to identify them case by case from some inconsistency of user profile (eg. a male user posts a lot about makeup; an actual case we found in the dataset) with the help of tools like Reddit User Analyzer [https://atomiks.github.io/reddit-user-analyser/] or SnoopSnoo [https://snoopsnoo.com/], but it would involve too much uncertainty and noise.

But still, there are considerations from the users that bot accounts disseminating negative content itself is not a great thing for the information ecosystem. In this sense, I think we can safely consider bots as "normal users".



We also tried to visualize moderator distribution for only top 100 commentators, and there are 80% moderators. Compared with the distribution of top 200 commentator, we think if continue looking at a wider range of commentator data, the percentage of moderator user would finally decrease. But we still think of the insight to be useful, because to a certain extent, it illustrates the correlation between being a moderator and being *actively* involved in negative discussions.

Finally, we plotted the **activeness** corresponding to user age. Activeness is reflected by total comment karma a user receives, and **user age** is inferred from register timestamp and newest activity timestamp. In the graph, the unit of user age is in days. There's not a clear distribution pattern, but it is easy to observe that moderators have a higher average of registration age and more instances of higher karma/age rate, while for non-moderators the increase of karma/age rate is significantly slower.

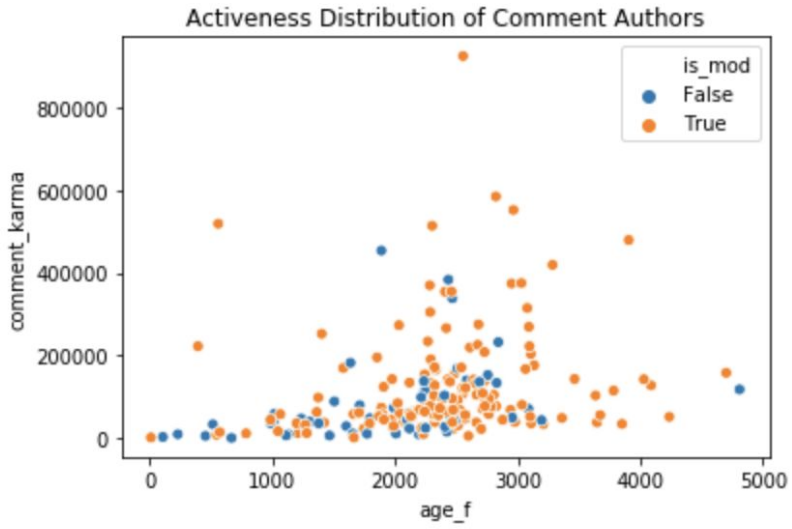


Figure 2-4: Distribution of activeness between moderator and non-mod

There might be an aggressive cancelling effect of upvotes/downvotes in comment karma for controversial users, so it is only a noisy approximation of activeness.

6. Reference

[1] S. Kumar, W.L. Hamilton, J. Leskovec, D. Jurafsky. Community Interaction and Conflict on the Web. World Wide Web Conference, 2018.

[2] Wardle, C. (2017, February 16). Fake News. It's Complicated.. First Draft.

[3] https://snap.stanford.edu/data/soc-RedditHyperlinks.html

[4] Perspective Api reference: https://github.com/conversationai/perspectiveapi/blob/master/api_reference.md

[5] PRAW API, https://praw.readthedocs.io/en/latest/

[6] nltk.sentiment.vader: https://www.nltk.org/_modules/nltk/sentiment/vader.html

7. Additional Link

[github] https://github.com/shakingkelly/TMD-reddit-conflict-analysis/blob/master/TMD.ipynb