Community Interactions On Reddit Over Time

Ava Trogus
atro7318@uw.edu

Daniel Fuchs
fuchsd@uw.edu

Aaron Burtle
amburtle@uw.edu

Josh Katz
katzjm@uw.edu

Abstract

Previous studies have explored the dynamics and implications of community interactions and conflict in online communities, but these studies often lack an analysis of the temporal trends in such interactions. In order to evaluate moderation policies and better understand long term platform viability, it's important to examine temporal trends in the absolute and proportional number of these conflicts. Numerically measuring these conflicts has typically been difficult since the criteria of a 'conflict' are highly subjective. In order to solve this, we utilize the findings of Kumar et al. to define and locate instances of interactions which are associated with "adverse effects and reduce the overall activity of users in the targeted communities" [1]. We then conduct a novel analysis on the temporal trends in these interactions. As a result, we demonstrate trends with implications towards the viability of the Reddit platform and synthesize a tool for measuring subreddit health and toxicity over time.

1 Introduction

Social media platforms are a ubiquitous and sometimes pervasive element of modern society. These platforms provide places for users from varying backgrounds to form communities and share content. However, with this sense of community also comes a sense of hostility. Indeed, hostile interactions within and between communities are commonplace enough to warrant the definition of terms such as trolling, brigading, and flaming. Considerable attempts have been made to examine the methods and implication of these hostile interactions as a whole [1]. However, as these social media platforms grow from their niche beginnings into multi-billion user services, it's important to examine not just the methods of these interactions, but how these interactions change over time. In particular, we examine how community interactions on Reddit are changing over time and investigate reasons for sudden changes in the dynamics of interactions towards particular communities.

We focus our analysis on Reddit for a variety of reasons. As a social media platform, Reddit allows users to create communities (subreddits) for sharing content in the form of text, hyperlinks, and images. However, unlike some social media platforms, Reddit users are exposed to content that is almost entirely generated by people that they have no real world connection to. As a result, communities on Reddit are less likely to be influenced by imposed social norms, instead gravitating towards chosen norms. This can lead to undesired results such as communities united behind racism and other anti-social ideologies. Additionally, Reddit users are pseudonymous. This allows users to provoke conflicts and use hateful language with minimal fear of real world consequences. Given these circumstances, an effective moderation policy is necessary on Reddit in order to discourage behaviours that lead to conflicts. Our methods provide quantitative measures of community conflict which can be used to evaluate and refine rules and moderation techniques. Additionally, we extend these measures to particular subreddits, both as a measure of subreddit health and as a tool to facilitate case studies.

Kumar et al. create methods for defining and locating community interactions which are associated with "adverse effects and reduce the overall activity of users in the targeted communities”. To facilitate our analysis we build upon these methods in order to locate hostile interactions and analyze trends in these interactions on a temporal basis. Based on personal experiences with Reddit, we hypothesize that community interactions will become more hostile with time.

2 Dataset

Reddit facilitates content sharing by allowing users to post content in communities of interest called subreddits. For example, www.reddit.com/r/politics is a community where users share news and opinions relating to American politics. Users on Reddit are also able to interact in various other ways such as comments and personal messages, but for the purpose of this study, we focus on interactions taking place in posts.

Drawing on the works of Kumar et al., we define community interactions as instances where a post on one subreddit links to a post on another subreddit. In order to get a list of such instances, we turn to Pushshift, a service dedicated to archiving Reddit posts and comments[2]. We use Pushshift instead of scraping data from Reddit directly for two reasons. First, the data is already formatted in first normal form, allowing us to easily utilize Spark for querying and manipulating the data. Second, historical data is kept by Pushshift. This allows us to analyze posts that have been deleted or removed by moderators some time after they were archived. However, since the data contains all posts instead of just posts containing crosslinks to other posts, additional processing is necessary to filter the dataset down to posts containing crosslinks.

Additionally, in order to provide a greater range for temporal analysis, we extend our methods to all Reddit posts originating between January 2014 and August 2019 instead of the January 2014 to April 2017 range used by Kumar et al. This window of time contains 1.5 terabytes of data consisting of 634,355,348 posts.

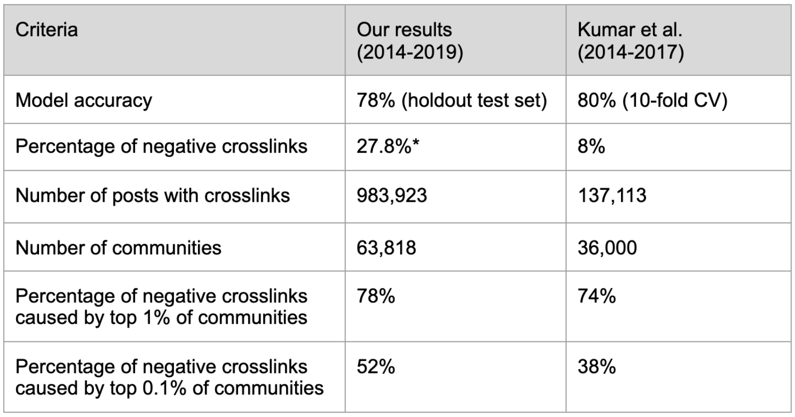
Before conducting our analysis, some additional data cleaning and filtering was necessary. Specifically, three inconsistencies in the data had to be resolved. First, the data type of the ’created_utc’ column had to be made consistent across all months of data. This required converting the column from a string into an integer for data occurring before 2015. Second, posts containing a null subreddit column had to be dealt with. We decided to remove these posts from our dataset after investigation indicated that these posts come from /r/Promos, a placeholder subreddit where Reddit advertisements are kept. Third, posts to subreddits with the prefix ’u_’ had to be removed. These posts correspond to a ’post-to-profile’ feature which allows users to create posts that reside on their own profile instead of being posted to any particular subreddit. These pseudo-subreddits do not reflect actual communities and were therefore removed [3].

Finally, for the purpose of training a sentiment model, we utilize all 1,024 posts and sentiment labels that were created by Kumar et al. for the same task. Specifically, we join the post ID of each row with the corresponding post in Pushshift. We then use the title stored in Pushshift in order to predict the sentiment of that post. Finally, we compare the predicted sentiment to the true sentiment as labelled by Kumar et al. in order to train our model. In order to provide a higher degree of confidence in our model, we labelled some additional data, resulting in 1,273 labelled posts in total.

3 Analytical Approach

Subreddit interactions are defined as events where users of one community observe, speculate about, or engage directly with users of another community. Drawing on the works of Kumar et al., we operationalize this construct by finding Reddit posts on one subreddit (the source community) that hyperlink to Reddit posts on another subreddit (the target community).

Interaction sentiment is defined as the general feeling of the source community towards the target community when an interaction occurs. Again drawing from Kumar et al., we operationalize this construct by analyzing the text associated with a post that links to another community and labelling that interaction as either ’negative’ or ’neutral’. This means that the ‘neutral’ category can be interpreted as non-negative, as it includes both neutral and positive posts. While it may seem arbitrary



Criteria	Our results (2014-2019)	Kumar et al. (2014-2017)
Model accuracy	78% (holdout test set)	80% (10-fold CV)
Percentage of negative crosslinks	27.8%*	8%
Number of posts with crosslinks	983,923	137,113
Number of communities	63,818	36,000
Percentage of negative crosslinks caused by top 1% of communities	78%	74%
Percentage of negative crosslinks caused by top 0.1% of communities	52%	38%

Figure 1: Model results converging with those of Kumar et al.

to stratify sentiment in this way, it is essential that we remain consistent with the methods of Kumar et al. in order to extend their results to our temporal approach.

To locate instances of community interactions through crosslinks, we began by running a regular expression query on the set of all 634,355,348 posts. As a result, we found 10,536,883 posts containing crosslinks to other subreddits. However, careful examination of these interactions showed that further data cleaning was necessary. Specifically, when examining crosslinks, we see that the title of the post in the source community is often just a copy of the title of the post that was linked to. Such cases where the title is copied violate our assumptions since the title presents no sentiment towards the target community. For this reason, we worked to eliminate such instances from our dataset.

To accomplish this task, we restricted crosslinks to instances where a post contains a crosslink to another post on a different subreddit. We then further compare the titles of the source and target posts via longest common subsequence. We filter out any posts where the longest common subsequence has a length greater than 90% of the target title, unless the source title is 1.5x longer than the target title. We use LCS instead of typical substring comparison in case of small copy paste errors and unicode encoding issues. Additionally, we allow for posts where the source title is significantly longer than the target title in case the target title is quoted but additional sentiment is appended to the quote. After filtering in this way, 983,923 posts containing crosslinks remain. It's worth noting that a large portion (approximately 7 million) of the posts filtered out were posts that link to non-post material in the target community (e.g. posts that link to comments). Given the lack of details on this subject in Kumar et al., we believe this additional filtering to be a novel addition to their original methods.

With community interactions extracted, we began the process of labelling interaction sentiment. Using similar methods to Kumar et al., we created a model to analyze and label the sentiment of a Reddit post with a crosslink. Our goal was to achieve a similar accuracy level as Kumar et al. through a different approach. They used a Random Forest classifier that considered many features of each source post. We, instead, trained a BERT model to analyze posts according to their word content. We tokenize the words in the 'title' section of each post, and run it through the model. The model weights words based on their association to the negative training data points, as well as the context of the other words in the post.

Our ground truth data comes from text data that was manually labelled with negative and neutral sentiment (1273 data points). This includes labelled data from the Kumar et al. study and some additional randomly sampled data that we manually labelled. We randomly shuffled the labelled data, and split it into an 80% training set, 10% validation set, and 10% test set. Once we tuned hyperparameters, we were able to achieve 78% accuracy on the hold-out test data. We then labelled the rest of our dataset with the trained model.

We measured the validity of our sentiment analysis model by comparing its performance to Kumar et al.'s classifier (see figure 1). As outlined in the comparison chart, their Random Forest classifier achieved 80% 10-fold cross validation accuracy which we were almost able to match with a 78% accuracy on the holdout test set achieved by our BERT model. One differing result we found was that our model labelled 27.8% of the total posts as negative, whereas the Kumar et al. model labelled only

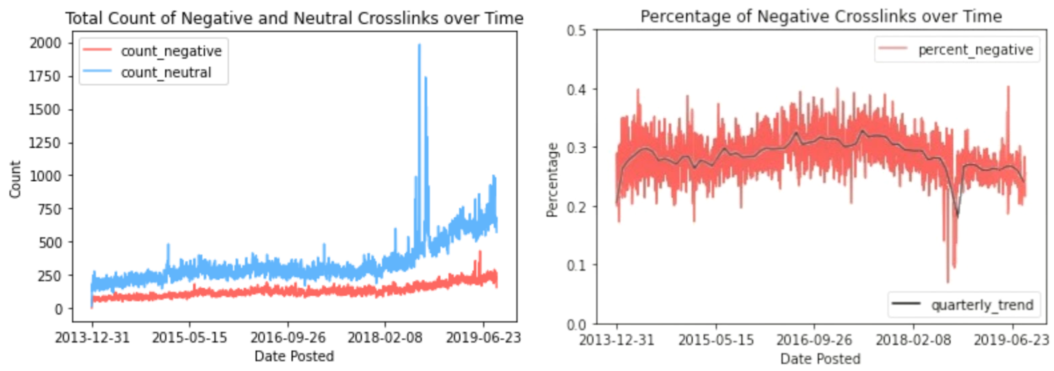


Figure 2: Aggregate trends on Reddit

8%. However, if we look at the training dataset, about 24% of those posts are of negative sentiment. Assuming a randomly sampled training dataset, our model is actually a better representation of the true distribution than the Kumar et al. model. We also found similar trends in that most negative posts are coming from the top 1% of subreddit communities and most of those negative posts are sourced from the top 0.1% of subreddit communities. Overall, our findings are convergent with the Kumar et al. study, even with significantly more data.

With the dataset now augmented with the labels from our sentiment analysis model, we now offer novel analyses of these Reddit posts with crosslinks according to the sentiment they carry.

First, we examine temporal trends in sentiment on the aggregate level. To accomplish this, we graph the number of negative and neutral sentiment posts as a function of time. We analyzed both absolute counts and relative proportions to get a full picture of the sentiment trends across the Reddit platform.

Second, we investigate particular subreddits of interest for notable spikes in interactivity during certain time periods. One area of interest was the sphere of political subreddits. We looked into a number of political subreddits such as /r/politics and /r/hillaryclinton. For each of these subreddits, we performed case studies by reading posts from time periods with significant increases in activity or changes in proportional sentiment. From these sampled posts, we tried to find a root cause for the sentiment change. We also looked at whether these increases were mostly characterized by negative or neutral sentiment to see how major current events impact subreddit culture.

4 Findings and Insights

4.1 Reddit on the Aggregate

First we looked at sentiment on Reddit as a whole. The left side of figure 2 (shown above) shows the result of taking the counts of all negative and neutral posts respectively and graphing them as a function of time. This graph demonstrates that Reddit has seen a gradual increase in both negative and neutral interactions over time. It is also apparent that the rate of increase grew in the beginning of 2018. This graph aligns with the growth of Reddit as a platform.

The large spike in neutral posts found in August and September of 2018 called for investigation. Our findings indicate that there were two major contributors to these spikes. The first was a coordinated attempt to facilitate free music sharing on Reddit via bot accounts. These bot accounts were regularly linking to posts by other bot accounts on different subreddits created for the same purpose. This generated a bunch of automated crosslinks, but eventually caused the communities and the bots to be banned. The second was a large volume of posts to /r/TranscribersOfReddit. We have decided to disregard these events as insignificant outliers.

The right side of figure 2 (shown above) tracks the proportion of negative crosslink posts over time. The day-to-day data varies so we also graphed a quarterly aggregate to more clearly analyze a trend. From 2014 until around the beginning of 2017, the data shows a slight upward trend in the proportion of crosslinks with negative sentiment, which supports our hypothesis. However, since then, this proportion has steadily dropped.

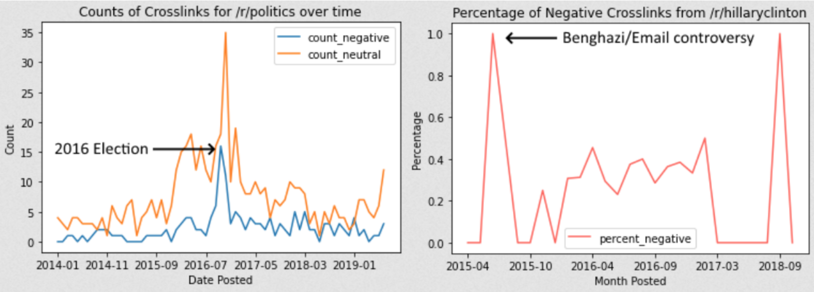


Figure 3: Controversial topics

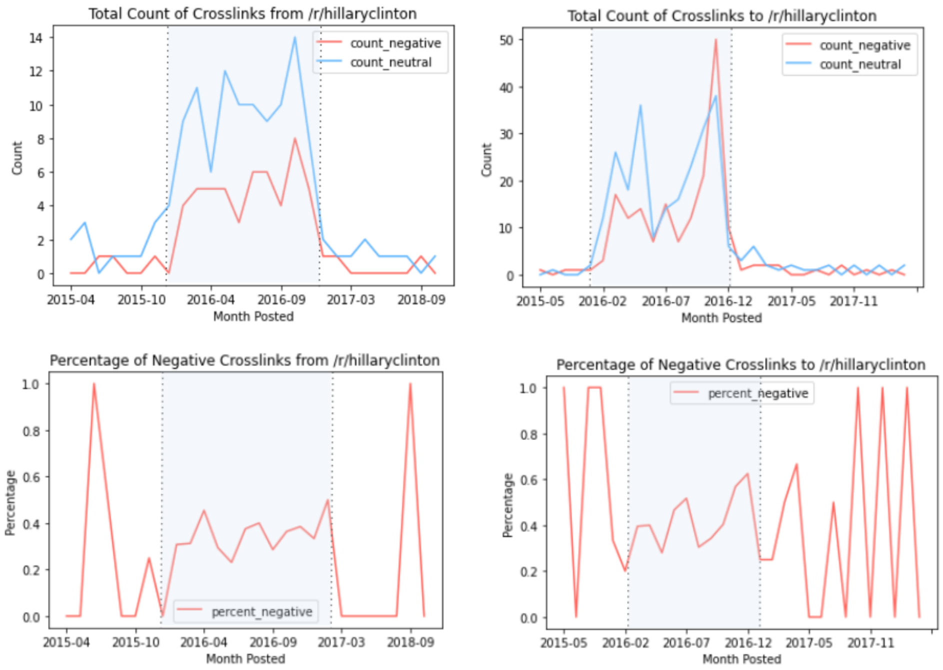


Figure 4: Crosslink changes over time into and out of /r/hillaryclinton

One potential explanation for this change in sentiment on Reddit is a series of changes to moderation policies. Not long after the proportion of negative sentiment crosslinks began to rise, a site-wide moderation policy was announced. These policies continued to be updated and refined until a dramatic incident occurred near the end of 2017. In that year, the CEO of Reddit, a user by the name of /u/spez, went so far as to "troll" /r/The_Donald by changing the contents of posts by users on that subreddit in a way that was intended to bother those particular users [4]. This action was taken as a direct result of /r/The_Donald's reputation as one of the most toxic communities on reddit. This is perhaps the most active that the leadership of Reddit has been toward specifically toxic subreddits, and we see a consistent downward trend in negative sentiments as a proportion of crosslinks following this event.

4.2 Political Subreddits

While the general trend on Reddit regarding the sentiments of crosslinks seems relatively stable with changes in the trends taking years to appear, certain types of subreddits appear to behave much differently. Within the political sphere of Reddit, the sentiments of crosslinks have much higher variance and trends can turn around in much shorter time periods. Furthermore, controversial real-world events seem to correlate strongly to changes in sentiments within these subreddits. For example, on the subreddit /r/politics, we can see a clear spike in both the count of negative sentiments, as well as the proportion of negative sentiments in the month prior to the 2016 election as shown in

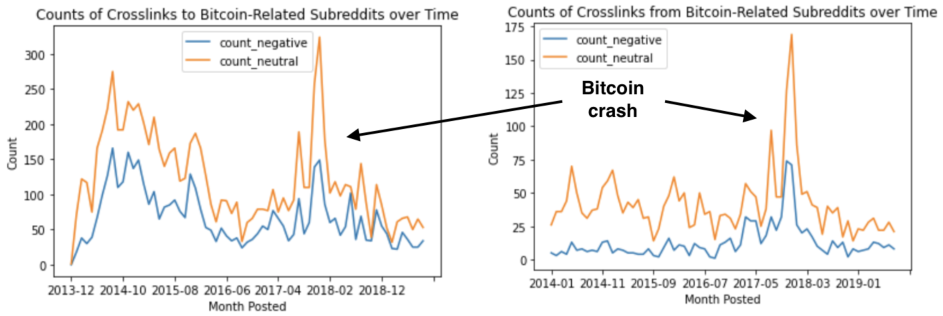


Figure 5: Crosslink Sentiments and the Bitcoin crash

figure 3 (shown above on previous page). This is then followed by a decrease in negative sentiments, but a clear increase in neutral sentiments in the month following the election as also can be seen in figure 3 (shown above on previous page). These correlations extend to other political subreddits as well, and can include more random, but still controversial events.

Also shown in figure 3 (shown above on previous page), we were able to observe a large increase in the percent of negative sentiment crosslinks from /r/hillaryclinton just following the link being made between the “Benghazi controversy” and Clinton’s “email scandal” in June of 2015.

We also discovered that sentiments can show bidirectionality, as is the case with /r/hillaryclinton and the 2016 election. Crosslinking activity increased not only from /r/hillaryclinton, but also also with respect to crosslinks targeting /r/hillaryclinton. These findings are shown in figure 4 (shown above on previous page).

As can also be seen in figure 4, both directions had similar negative sentiment rates of around 50%, which shows a significant increase in hostility in relation to the general dataset’s negativity rate of 28%.

4.3 Bitcoin Subreddits

Another big reaction to a current event that can be observed in bidirectional Reddit controversy is the 2018 bitcoin market crash. From January to February of 2018, the value of bitcoin dropped 65%. As can be seen in figure 5 (shown above), we grouped the top 3 bitcoin-related subreddits (/r/bitcoin, /r/btc, and /r/BitcoinBeginners) and graphed the total counts of negative and neutral crosslink posts targeting and sourced from these subreddits. We can observe in figure 5 (shown above) that in the early months of 2018 there is a huge uptick in interactivity in both scenarios. And as this graph shows, we can see the highest volume of posts with negative sentiment crosslinked in both directions.

5 Limitations

While our findings correlate with real world events and are convergent with the works of Kumar et al., our methods still face some limitations. In the particular case of construct validity, our model faces some challenges. In order to further validate our model, we attempted to stratify subreddits into sarcastic and non-sarcastic subreddits. We then sampled some data from each and compared our model performance in both cases. We saw a 95% test accuracy among non-sarcastic subreddits, but only a 64% test accuracy among sarcastic subreddits. This indicates that further work is necessary in order to correct systematic bias that our model may introduce.

We also have some limitations with respect to internal validity. In particular, we only examine Reddit posts that link to other Reddit posts. However, the volume of user participation is much greater in Reddit comments. In order to verify these results, it would be good to reapply these methods to Reddit comments and ensure that the same correlations are found.

Additionally, while these methods have implications to social media platforms other than Reddit, our work clearly has limitations in this regard as well. Reddit is effectively an anonymous platform, allowing users to create as many new accounts as they would like, with little to no confirmation of identity. Platforms such as Facebook, where users are more closely tied to a real world persona would have different conditions under which interactions take place. Therefore, findings should be carefully validated when applied to different domains.

Lastly, it seems that community norms and moderation policies also have a strong effect on how sentiments are displayed on community-based websites, and so this also must be taken into consideration when examining our conclusions within the context of another online community.

6 Related Work

6.1 Community Interaction and Conflict on the Web

Much of our analysis was built upon the work of Kumar et al. The idea of understanding Reddit community interactions through crosslinks came from this paper, as well as the idea of labelling each of these posts according to their sentiment towards the post they are linking to. Kumar et al. created a model that labelled crosslink sentiments with 80% accuracy, which then became the target for our own sentiment model. The purpose of Kumar et al., however, was to find a pattern in how crosslinks mobilizing one subreddit's users onto another (either in negative or neutral contexts) changed the subreddits involved. They found that less than 1% of communities on Reddit start 74% of the conflicts. Additionally, they found that when communities mobilize against another subreddit, the target subreddit tends to lose members and some of the colonizing users stay on the target subreddit, but that active engagement by the colonized subreddit against the mobilization reduces the impact. [1]

6.2 Extracting Inter-community Conflicts in Reddit (Datta et al)

Datta et al. created a directed graph with weighted edges to illustrate a conflict network between subreddit communities. Each edge weight represented the magnitude of conflict between subreddits and each edge was directed from the source of the negativity to the target. A conflict was defined by the number of authors that are "social" (most posts receiving upvotes) in one subreddit and "anti-social" (most posts receiving downvotes) in another subreddit. This paper inspired us to analyze whether negativity spikes from a particular subreddit correlate with negativity spikes toward that same subreddit. [5]

6.3 A Large Self-Annotated Corpus for Sarcasm (Khodak et al)

Part of validation was to determine how our sentiment model performs on sarcastic versus regular data. To assist, we used data from this sarcasm dataset, which is an unbalanced compilation of all of the sarcastic comments on Reddit from 2009 to 2016. The comments were self labelled with the "/s" annotation, which comment authors used to explicitly mark their comment as sarcastic. With this dataset we were able to compile a list of the most and least sarcastic subreddits for further validation of our model, as well as get a rough idea of how sarcastic Reddit is as a whole and how much this might have affected our results. [6]

7 Conclusion

The sentiments of crosslinks on Reddit offer an insight into how the various communities there, known as subreddits, view one another, and provide a way to measure the hostility between one another. Too much of an increase in inter-community hostility may drive away users that wish to enjoy an environment free from toxicity. Measuring this inter-community hostility offers a way to predict the viability and health of online community-based platforms. Unsurprisingly, controversial real-world events play a role in influencing the observable levels of hostility from one community toward another. The ability to measure how these events correlate to changes between the hostility levels of different communities offers an insight into how moderation techniques can be more efficiently applied to those cross community interactions which will most need it. Furthermore, measuring the overall level of hostility between communities on platforms such as Reddit allows for the analysis of moderation changes, and changes in community norms, in order to recognize when these changes are having a positive impact on the platform as whole.

It appears that the changes which have taken place on Reddit since 2017 have led to fewer negative cross community interactions as a proportion of total cross community interactions (as measured from crosslink sentiment analysis), and yet a clear upward trend of overall negative sentiments can be seen. While our prediction that negative sentiments would increase on Reddit over time does seem to have come true, and that negative sentiments as a proportion of cross links did go up from 2014 until 2017, it remains unclear how the recent upward trend of neutral sentiments vastly outpacing negative sentiments, and therefore causing negative sentiments to be a smaller proportion of crosslinks, will effect the long term sustainability of Reddit. Will the increase in total negative sentiments be washed away from the abundance of neutral sentiments, or will the added total of negative sentiments outweigh the reduction in negative sentiments as a proportion? More time and research is needed to tell, but what is clear, is that we are likely to see a spike in negative sentiments within certain communities as controversial real world events invade the communities that we form across the internet.

References

[1] Srijan Kumar, William L. Hamilton, Jure Leskovec, and Dan Jurafsky. Community interaction and conflict on the web. *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW'18*, 2018.

[2] *Pushshift API*. `https://pushshift.io/`.

[3] Daniel Fuchs, Ava Trogus, Aaron Burtle, and Josh Katz. *Project Repository*. `https://gitlab.cs.washington.edu/fuchsd/cse481ds-reddit`.

[4] Steve Huffman. *TIFU by editing some comments and creating an unnecessary controversy.*, 2016. `https://www.reddit.com/r/announcements/comments/5frg1n/tifu_by_editing_some_comments_and_creating_an/`.

[5] Srayan Datta and Eytan Adar. Extracting inter-community conflicts in reddit, 2018.

[6] Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. A large self-annotated corpus for sarcasm, 2018.