Deplatforming and Intra-platform User Migration: A Case Study

Course Project for CS598 Antisocial Computing - FA21

ALLEN ZHANG, University of Illinois at Urbana Champaign
JERRY NIE, University of Illinois at Urbana Champaign
GitHub Repository Link: https://github.com/AllenJRZhang/CS598-Antisocial

1 Introduction

When dealing with online social platforms, especially content moderation, "Deplatforming" is one concept that cannot be avoided. As major platforms such as Reddit or Twitter are paying more and more attention to content moderation in recent years, deplatforming is also happening much more frequently than ever. With its importance growing from both user and researcher perspective, we feel the need to examine this concept, what it means, why it's important and its impact a little closer.

1.1 What is deplatforming

First, we must answer the question: What is deplatforming? The concept of 'Deplatforming' refers to "the permanent ban of controversial public figures' from some online platform. (Jhaver et al., 2021) Basically, when a user's behavior is deemed as unacceptable or offensive, he/she is blocked, or 'deplatformed' from the platform and prevented from further participation on that platform. In addition to individual users, platforms have the ability to 'deplatform' an entire community or group.

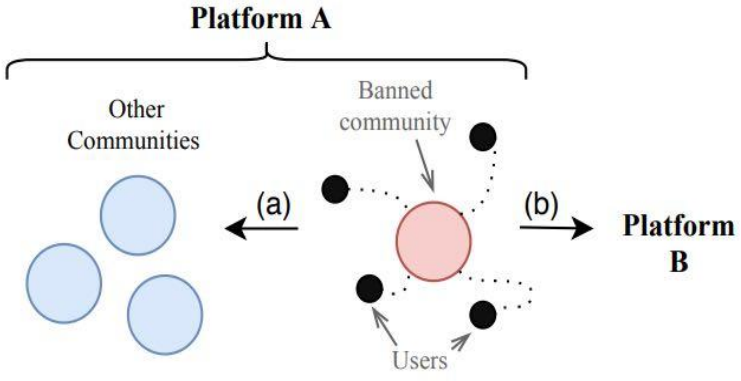

Figure 1. Illustration of Deplatforming and User Migration

1.2 Why deplatforming may cause potential problems?

Then, we ask ourselves why may deplatforming cause potential problems? Indeed, from a platform perspective, deplatforming as a method of content moderation is effective; as the most toxic users get banned, the environment of the platform itself would normally become healthier. However, deplatformed users normally just migrated to other platforms, thus the overall toxicity level of the entire Internet ecosystem has not decreased.

Therefore, we intend to use scientific approach, to conduct a case study, and discover the impact of deplatforming and user migration on the overall toxicity level.

Since related literature has already covered the scenario when users migrate from platform A to platform B following the deplatforming happening on the former, we are more curious about Intra-platform migration, namely, users migrating from the banned community to other communities of the same platform.

2 Research question

The main research question for our study is “How would deplatforming and intra-platform user migration affect the overall toxicity level of a certain community?” As we’ve stated earlier, we focus on intra-platform user migration. (i.e., migrating to another sub-community of the same platform) To answer this question, we will conduct a case study on *r/The_Donald* and *r/BannedFromTheDonald*, two subreddits which involve the behavior of deplatforming and user migration.

We expect three different potential outcomes: 1) Toxicity decreases. One possible explanation is that users may have learned to behave themselves after getting banned from the previous community. 2) Toxicity increases. One possible explanation is that users may get even angrier because of the ban and subsequently become more toxic in the new community. 3) Toxicity remains the same. It could be the case that users maintain the same behavior, just switching to a different place.

3 Related works

In the paper “Do Platform Migrations Compromise Content Moderation? Evidence from r/The_Donald and r/Incels”, the authors researched what will happen to the community after their forum is moderated. They examined r/The_Donald and r/Incels and their standalone websites like thedonald.win and incels.co. According to their analysis, the community-level moderation measures decrease the capacity of toxic communities to retain their activity levels and attract new members, which proves the moderation is effective.

As for the paper “Community Interaction and Conflict on the Web”, mainly focused on user interactions within different subreddits. In this paper, the authors believe every sub reddit could be regarded as an online community. These communities not only provide a gathering place for intra-community interactions between members of the same community, they also facilitate intercommunity interactions, where members of one community engage with members of another. They came up with a concept called mobilization. They define mobilizations as cases where a cross-link leads to an increase in the number of comments by current source members on the discussion thread of the target post. They showed that a reduction in the echo-chamber effect, and an increase in defenders’ use of angry words towards attackers, are associated with decreased rates of colonization and increased future participation rates for the defenders. Thus, it appears that increased engagement and a more fierce defense may be a more effective mitigation strategy, compared to ignoring or isolating the attacking users.

4 Data collection

We have collected data from three subreddits we are interested in: *r/The_Donald*, *r/BannedFromTheDonald* and *r/donaldtrump*. *r/The_Donald* is the community deplatformed users migrate from, and *r/BannedFromTheDonald* is the community deplatformed users migrate to. Whereas *r/donaldtrump* mainly serves as a reference group.

We used Reddit API to collect data from *r/BannedFromTheDonald*. We have collected 20066 comments from 20000 posts. However, since the Reddit API can not provide data assessment for

a banned subreddit such as *r/The_Donald* and *r/donaldtrump*, we used the Pushshift Reddit dataset to collect data from these two subreddits. We randomly selected 20,000 comments from *r/donaldtrump* as well as 20,000 comments from *r/The_Donald*. We use these 40,000 comment data for sentiment analysis.

In order to analyze the user migration, we also collected the user data of these three subreddit. For *r/The_Donald*, we collected user information data about all users who had commented on it in the two years before it was banned, including 16.4 million comments from 271, 102 commenters. For *r/BannedFromTheDonald*, we collected all the comments in the three years before December 15, 2021, and 9,138 commenters posted a total of 54,615 comments.

5 Methods and techniques

After collecting enough data, we will analyze and compare the data. The study consists of two types of analysis: user-level analysis and community-level analysis. Details of each are shown as follows:

5.1 User-level analysis

To better understand changes at the user-level, the first thing we want to study is how many users of the original r/The_Donald/ have immigrated to these two subreddits that we want to study. This can be achieved by comparing the users of r/The_Donald/ with the users of related subreddits.

One problem is the pushshift api does not provide functionality for retrieving all subscribers of a certain subreddit. In order to retrieve users/subscribers of a subreddit, all we must do is collect as many comments as possible and record their authors. The first thing we need to determine is the time span of the comments that will be collected. After determining the time span, start to collect comments, and exclude users whose authors are automoderator or have been deleted (marked as [deleted] by pushshift), count these commenters and use a dictionary to save the commenters that have been recorded.

Since the scale of users of different subreddit is very different, we have not been able to collect all users' data. For *r/The_Donald*, we collected user information data about who had commented on it from June 6th, 2018 to June 5th, 2020 (banned on June 29th, 2020), including 16, 400, 938 comments and 271, 102 commenters. For *r/BannedFromTheDonald*, we collected all the comments in the last three years (from December 15th, 2018 to December 15th, 2021), and 9,138 commenters posted a total of 54, 615 comments.

5.2 Community-level analysis

Now that we have user-level analysis to support the significance of choosing the selected communities, we can conduct community-level analysis on them. The community-level analysis consists of three parts: sentiment analysis, top 50 most frequent positive/negative words and, most importantly, toxicity level analysis.

To conduct any of these analyses we must first preprocess the data. We chose to use a unigram language model for the analyses, we need to transform our data from the string structure representing individual comments to word-level representation. We do this by adopting the python NLTK package. First, we remove all the emojis from our input data. Then, we tokenize the input and convert everything to lowercase. After that, we remove all the stop words (words that are too common to be of significance, such as “the”, “a”, “he” and “she”). Last but not least, we use both stemming and lemmatization techniques to make the whole data collection more robust and concise.

Now that we have the preprocessed data at our disposal, we can conduct the analyzes we want. The first is sentiment analysis.

For sentiment analysis, we utilize a pre-trained sentiment classifier from the NLTK package, and apply it on our now-preprocessed data. The data is then divided into three sentiment categories: positive, negative and neutral. Since naturally a large chunk of the data would be categorized as neutral, which is not significant for our study and it makes the other two groups less obvious, we exclude the neutral category, keeping only positive and negative results for comparison. Finally, we use a bar chart for data illustration.

Now we have all the important words categorized either as positive or negative, we can sort each group by frequency in the dataset to generate the top 50 most frequent words in either positive or negative categories. We use a histogram to represent the frequency of these words. By observing this result, especially the negative words, and comparing them across communities, we can get our first glimpse of toxicity level. Namely, if a given community has a relatively high level of toxicity, we are expected to see negative words that are either offensive, obscene, or aggressive appearing more frequently. On the other hand, if a community's toxicity level is relatively low, then we expect the most frequent negative words for this community to be more generic and less offensive, such as "dislike", "bad" or "awful".

Next, we directly look at the toxicity levels of our chosen communities. Similar to sentiment analysis, we do this by applying a pre-trained word toxicity classifier on our input data. And it will give a score for each of the six toxicity categories for a community: *overall toxicity, severe toxicity, obscene, threat, insult* and *identity attack*. We can now compare the toxicity level of two communities by observing and comparing the toxicity scores we have generated.

6 Challenges and problems

In the process of conducting this research, we inevitably met certain challenges and obstacles. The following paragraphs detail several challenges we encountered during the process of completing this study.

The first problem is the Pushshift api has some limitations. Pushshift does not record the data on reddit in real time, so the data of pushshift is not necessarily complete, and there may contain small-scale data loss. Also, pushshift does not provide an api for retrieving all subscribers of certain subreddits, which means that the subscribers of a certain subreddit can only be obtained by retrieving comments and then recording the comments' authors.

The second problem is computer performance constraints. As we all know, reddit is one of the hottest forums in the world. Even a banned suredit contains millions of data. After we finished

our code and conducted a small scale test, we noticed that we may not be able to analyze the whole subreddit. Retrieving millions of comments is a heavy burden. We decided to randomly select 20,000 comments from two different subreddits to apply our sentiment analysis. The computer machine we used is a Linux workstation laptop, equipped with AMD R7-4800H and 32 Gigabyte memory. Even if there were only 40000 comments and the data were analyzed by batch, the execution time of our program exceeded 12 hours.

We also face the limitation of the concept of case study. This is only one specific case of deplatforming and intra-platform user migration that we are studying. It is hard to say if the results we've found can also be applied to other similar instances or would different cases lead to different, if not contradictory results. Further studies of broader scale are required.

In addition, what we used for sentiment and toxicity analysis is a unigram language model. The sentences and paragraphs are broken down into individual words. However, both sentiment and toxicity analyzes are heavily dependent on the semantic content of the input, therefore using a unigram language model means losing some significant information. For example, the toxicity level of a sentence is not necessarily the sum of the toxicity levels of each word in that sentence. Thus, our results only serve as a good approximation of the ideal result. To make the analysis more accurate, we need more complex language models and algorithms.

We also didn't consider the context information of the data we collected. Different comments should be analyzed in their own context, but the reality is that some comments lack context, so we chose to aggregate all comments together and do the analysis without considering context information, which means we lost some valuable information.

Last but not least, we also need to consider the difference between correlation and causation. We did find a correlation of deplatforming, user migration behavior and the toxicity level. However, such a correlation does not imply one causes another. Many other factors may also contribute to this matter. Some additional techniques would perform better in exploring a potential causal relationship between the behavior that we are interested in and the toxicity level of a given community.

7 Results and interpretations

Now we illustrate the results and findings of our study and interpret them accordingly.

7.1 User-level analysis

Since *r/BannedFromTheDonald* is identified as the subreddit where mostly only banned users from *r/The_Donald* gather, ideally the user overlapped ratio between the two communities should be 100%. However, this naive expectation cannot be reached in reality since the limitation of PushShift API only allows us to retrieve the username of users making comments in a certain period of time, not all users from either subreddits can be accounted for.

Since users were migrated to *r/BannedFromTheDonald* after being banned by *r/The_Donald*, there will be a certain lag in the data. Among the 9, 138 users who have posted comments on *r/BannedFromTheDonald* in the past three years, 1, 137 of them have commented in the last year of *r/The_Donald*. If we change the time span from one year to two years, this number increases to 2, 351. Due to the performance limitation of the PushShift API (approximately 48 hours of running time per 10 million comments), we cannot collect more data to prove that most users of *r/BannedFromTheDonald* are from *r/The_Donald*. Intuitively though, if we are able to collect more data from a longer time range, we expect the percentage of user overlapping to continue to increase. Nonetheless, the user-level analysis did provide evidence that the user base of these two communities are correlated and somewhat overlapped, which is enough for the community-level analysis to be meaningful.

7.2 Community-level analysis

First we have the results from sentiment analysis.

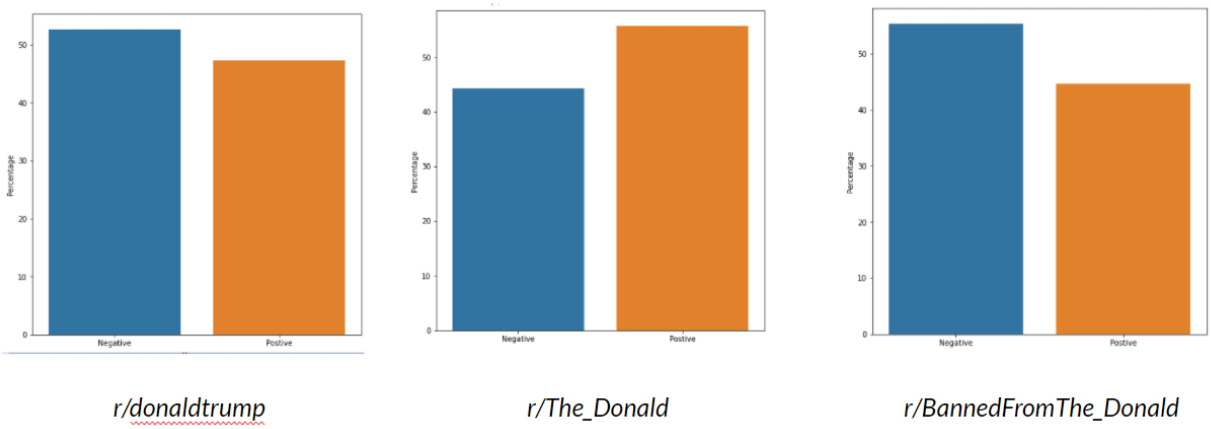

Figure 2. Sentiment analysis results (excluding neutral)

As we can see on the above illustration, the sentiment patterns for *r/The_Donald* and *r/BannedFromTheDonald* are almost inverse of each other, whereas *r/donaldtrump* follows a similar pattern as *r/BannedFromTheDonald*. The overall sentiment of *r/BannedFromTheDonald* is more negative than *r/The_Donald*. Two possible factors may contribute to this result. The first is that it is very likely that the deplatforming of users migrated from *r/The_Donald* to *r/BannedFromTheDonald* affects their sentiment, which is skewed towards negative. Also, it may have something to do with the difference of main topics being discussed in each subreddit. *r/The_Donald* mainly discussed the campaign and policies of former president Donald Trump, who, despite all the controversy, did have a large number of supporters active in that subreddit, which is indicated by the larger percentage of positive sentiment of *r/The_Donald*. However, on the other hand, *r/BannedFromTheDonald* mainly consists of banned users from *r/The_Donald*. Main topics in this subreddit might involve their bans and critiques of Donald Trump more often compared to *r/The_Donald*, which leads to the opposite result of sentiment analysis. In order to further solidify this interpretation, we need to conduct additional topic mining studies.

Next we have the top 50 most frequent positive/negative words for each community. First we show the comparison of positive category across all three subreddits:

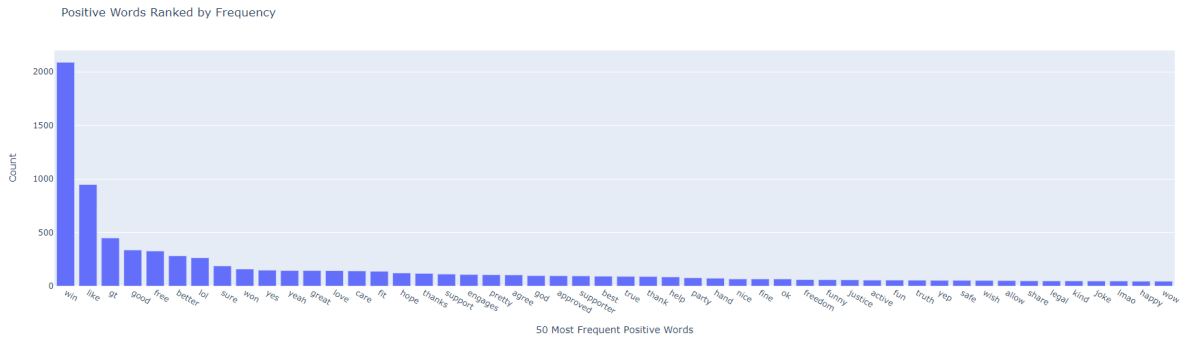

Fig 3: Top 50 most frequent positive words for *r/The_Donald*

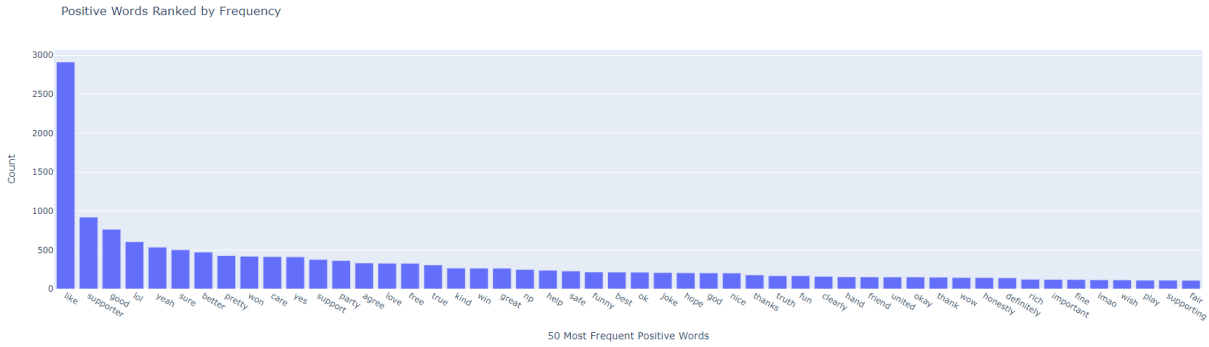

Fig 4: Top 50 most frequent positive words for *r/BannedFromTheDonald*

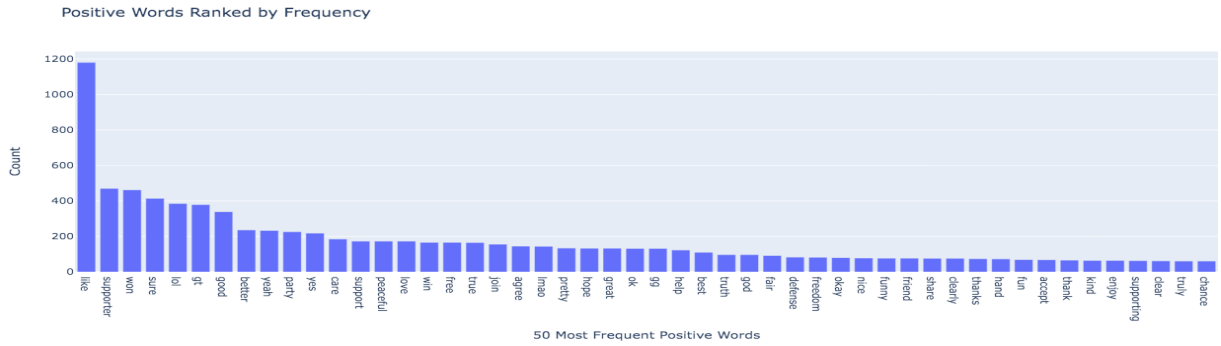

Fig 5: Top 50 most frequent positive words for *r/donaldtrump*

As we can observe from the above figures, the positive word distributions are very similar across three communities, meaning there's not a distintable difference in positive sentiment among these subreddits. Therefore its significance is trivial as we are more interested in the negative word distribution, which is more correlated with our main focus: toxicity:

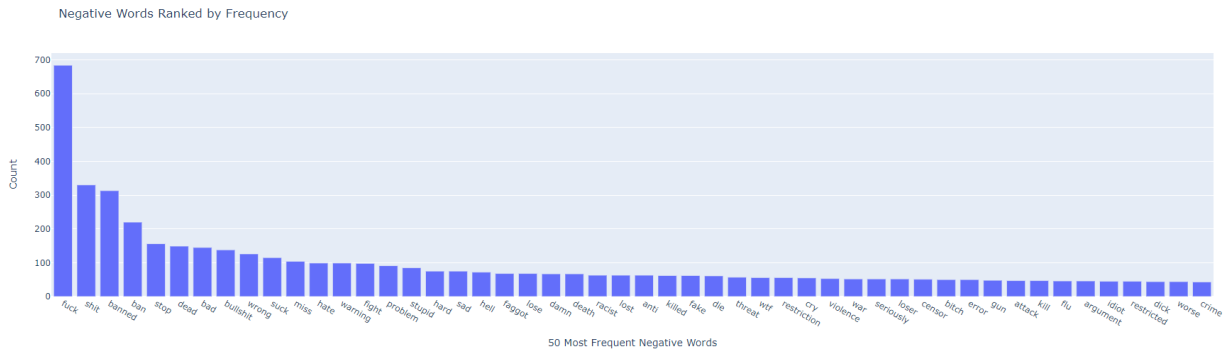

Fig 6: Top 50 most negative frequent words for *r/The_Donald*

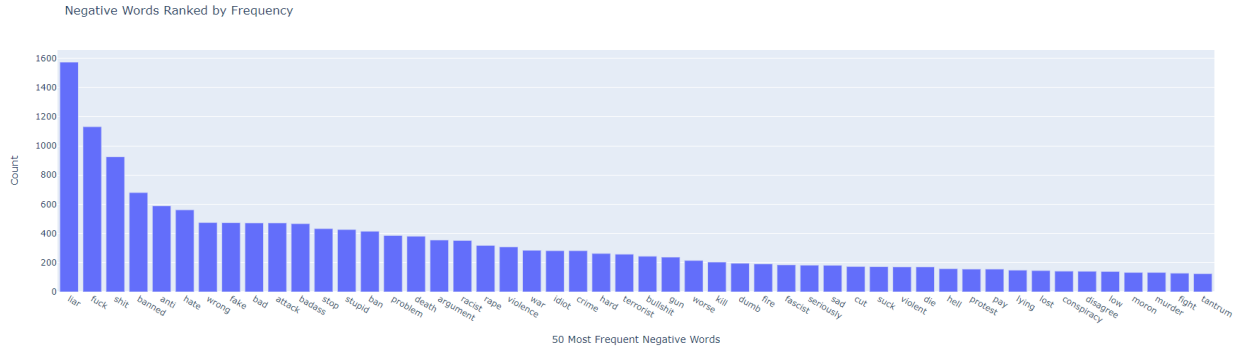

Fig 7: Top 50 most frequent negative words for *r/BannedFromTheDonald*

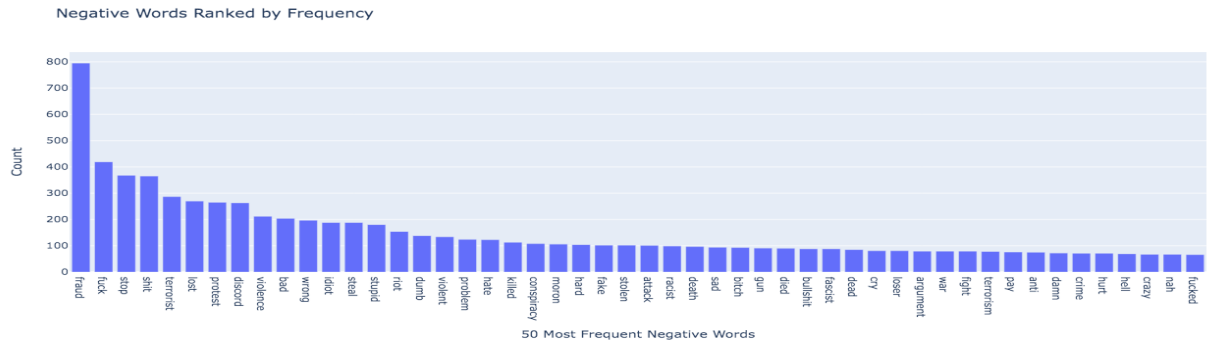

Fig 8: Top 50 most frequent negative words for *r/donaldtrump*

From the above word distributions and patterns, we can tell that the most frequent negative words of all three subreddits are all very toxic, with *r/The_Donald* and *r/BannedFromTheDonald* slightly more offensive than *r/donaldtrump*. Also, in *r/The_Donald*, the word “fuck” is used way more frequently than any other negative words, wheras in *r/BannedFromTheDonald*, many toxic words have relatively high frequency. This implies that *r/BannedFromTheDonald* is potentially more toxic overall than *r/The_Donald*. However, both subreddits share strikingly similar word

patterns, which further suggests that the user bases of these two communities are indeed correlated, or even overlapped.

Finally, we have the results for our toxicity level analysis:

	r/donaldtrump	*r/The_Donald*	*r/BannedFromThe_Donald*
Toxicity:	0.04792614449635432	0.04996599619881574	0.0620726573811539
Severe toxicity:	0.00289363749564698	0.00402696861451814	0.0035699851985770137
Obscene:	0.018258539530737347	0.02482314147278014	0.02316678586603872
Threat:	0.002031542501911648	0.0019901723509775917	0.0033824538646196074
Insult:	0.012095015534366638	0.012376988745004563	0.014467081988492426
Identity attack:	0.0019067082356493028	0.001877537290213373	0.0033096089623717534

Fig 9: Results for toxicity level analysis

As we can see, *r/BannedFromTheDonald* has the highest overall toxicity, whereas *r/The_Donald* has the highest score of severe toxicity. *r/BannedFromTheDonald* and *r/The_Donald* are roughly equal on obscenity level. *r/BannedFromTheDonald* also has the highest score on the rest three categories. In summary, *r/BannedFromTheDonald* is overall more toxic, but *r/The_Donald* has some extreme toxic cases. One possible explanation is that deplatformed users might be infuriated because of their ban. And after migrating to another community, they tend to become more toxic as a way of venting their anger. This suggests that although banning is one of the most effective moderation approaches, it only works for a particular community. It is not necessarily able to decrease the overall toxicity of an entire platform, or even the Internet as a whole.

8 Directions for future study

One potential direction for future study is to conduct topic mining on our selected communities. By analyzing the most discussed topics of each community, we will be able to provide evidence either supporting or countering our interpretation for the result of sentiment analysis. That is, the difference in sentiment between two communities is likely due to the difference in main topics being discussed in each community.

Also, since we already discussed the limitation of using a unigram language model, to make our results more accurate and more meaningful, it is a good idea to use more complex language models and algorithms to maintain the semantic information of input comment data. Context data can also be taken into account to make the analysis more robust.

Last but not least, this is only a case study, so similar research can be done on other instances of deplatforming and intra-platform user migration to see if our results can conform or contradict with other cases. Eventually, we hope to generalize the findings to deplatforming and user migration as a whole, instead of individual cases.

References

Ribeiro, M.H., Jhaver, S., Zannettou, S., Blackburn, J., De Cristofaro, E., Stringhini, G., & West, R. (2020). Does Platform Migration Compromise Content Moderation? Evidence from r/The_Donald and r/Incels. ArXiv, abs/2010.10397.

Jhaver, S., Boylston, C., Yang, D., & Bruckman, A. (2021). Evaluating the Effectiveness of Deplatforming as a Moderation Strategy on Twitter. Proc. ACM Hum.-Comput. Interact.

Zhang, J. S., Keegan, B. C., Lv, Q., & Tan, C. (2020). Understanding the Diverging User Trajectories in Highly-related Online Communities during the COVID-19 Pandemic. arXiv preprint arXiv:2006.04816.

Kumar, S., Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2018, April). Community interaction and conflict on the web. In Proceedings of the 2018 world wide web conference (pp. 933-943).

Cho, S. Y., & Wright, J. (2019, November). Into the Dark: A Case Study of Banned Darknet Drug Forums. In International Conference on Social Informatics (pp. 109-127). Springer, Cham.

https://en.wikipedia.org/wiki/Controversial_Reddit_communities