Challenge 2 : Moderation for Reddit

1. Problem description

As a data scientist at Reddit, our task is to suggest new ways to alert the community manager of inappropriate behavior on social media. Indeed, in view of the Digital Service Act regulation of October 19, 2022, what is illegal offline is also considered illegal online. Inappropriate content on Reddit's platform should then be regulated.

To tackle this challenge, we will engage in a comprehensive analysis that consists of evaluating user behavior through toxicity scores provided by BERT (Bidirectional Encoder Representations from Transformers) and analyzing community interactions through graph theory. Our approach involves representing interactions between users as an undirected and weighted graph where nodes represent users and edges represent interactions such as comments and replies, enabling us to detect patterns and clusters indicative of toxic subreddits. This dual analysis will enable us to provide the client with prioritized lists of users and subreddits that require enhanced monitoring, ensuring compliance with the DSA and fostering a safer online environment. The expected outcomes of this project include offering recommendations to Reddit's community management teams for necessary interventions.

2. Data analysis

The datasets provided for this work are centered exclusively on climate-related discussions on Reddit. We have four key files that give us a comprehensive view of how climate topics are discussed within the platform. The first file lists various subreddits dedicated to climate issues, providing a broad view of the communities engaged in environmental discussions. The second file narrows this list down to those subreddits that are currently active. The third file contains details of all posts made in these active climate-focused subreddits, offering insights into the specific topics and types of content that engage users. Finally, the fourth file includes all comments from these posts, allowing us to delve deeper into the nature of the discussions and the community's responses.

The initial analysis of the data from active climate-focused subreddits has provided valuable insights into user engagement and content dynamics on Reddit. Overall, we have data on 64 subreddits, 24 of them being considered active, which reduces the number of subreddits we have to concentrate moderation on. The 'all_posts_active_subreddit.csv' dataset reveals that certain subreddits, such as 'climateskeptics', 'climatechange', and 'ClimateOffensive', are particularly active, indicating high levels of user interaction and discussion intensity. These forums, alongside others listed in the top ten, are vital areas for targeted moderation, especially under the stringent requirements of the Digital Services Act, which mandates proactive measures against harmful content. The comment data analysis from 'all_comments.csv' shows that there are 337817 comments in our dataset, this has to be taken into account when using models to process our data as it can become computationally expensive. Comments are in different languages including English, German and French. Therefore, we have to be able to detect toxicity in different languages.

Going further, we can study when peak activity occurs on the platform each day. This could help us understand when it is important to concentrate moderation efforts.

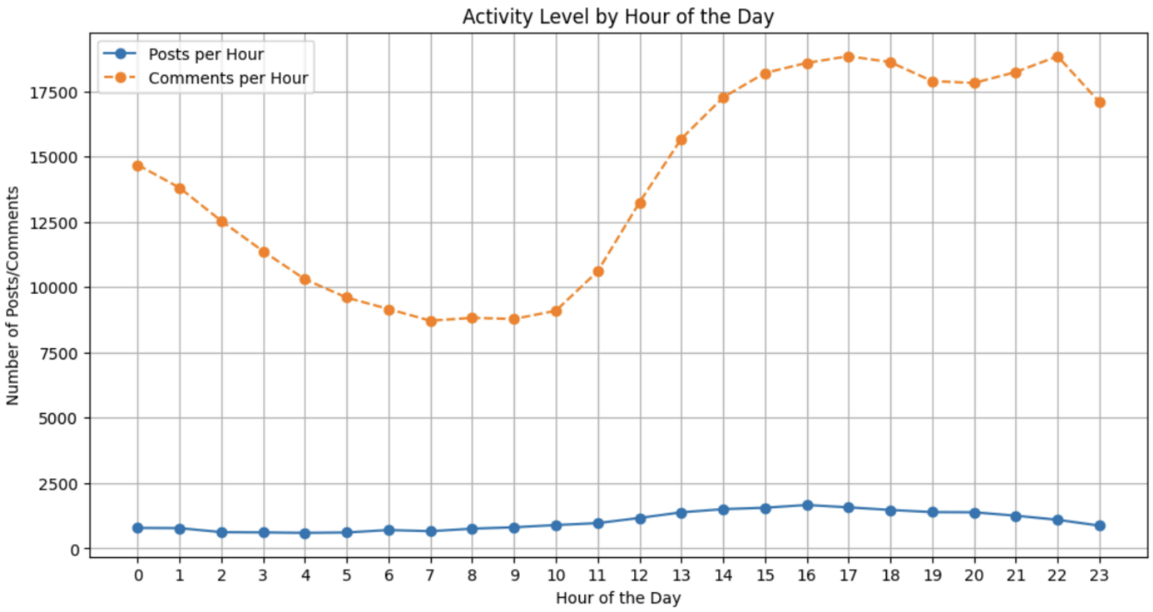


Figure 1 : Activity level by hour of the day

The hourly activity analysis of Reddit posts and comments reveals distinct patterns: peak activity occurs at 16:00 for posts and at 17:00 for comments, suggesting a trend where posts likely prompt responses that peak in the subsequent hour. The activity for both comments and posts gradually increases from morning, stabilizes through the afternoon, and evening, and then sharply declines after 22:00, indicating minimal interaction during late night and early morning hours. These insights are important for optimizing moderation strategies, as they suggest the need for heightened moderation during peak hours to efficiently manage user interactions and maintain community guidelines.

3. Toxicity scores

BERT, which stands for Bidirectional Encoder Representations from Transformers, is a model of natural language processing (NLP) developed by Google. This model uses bidirectional representations from text, making it able to learn a rich understanding of language. For the purpose of moderating online platforms like Reddit, BERT can be particularly useful in assessing toxicity in user-generated content. By fine-tuning BERT on a dataset of labeled comments—where texts are marked as toxic or non-toxic—the model can learn to predict the likelihood that a new, unseen comment is toxic. This capability allows community managers to automatically flag potentially harmful content for review, supporting efforts to maintain a respectful and safe online environment. This application not only enhances the efficiency of moderation but also helps in adhering to digital safety regulations such as the Digital Services Act.

As we did not have labeled data given, we decided to work with a pre-trained Bert model. Indeed, it would have been too time consuming to label each comment by ourselves within the scope of this project. Also, in view of the high number of comments that we have, training a model specifically on it would take a very long time. Therefore we decided to work with the model “unitary/toxic-bert”. This model is trained on Jigsaw dataset, which is a collection of online comments featuring comments labeled for various forms of inappropriate content, including toxicity, severe toxicity, threats, insults, and identity-based hate. As said before, the comments of our dataset are in various languages. “Unitary/toxic-bert” is a multilingual model working with 7 languages. It returns a score between 0 to 1 for each comment with 1 meaning that the comment is very likely to be toxic, 0 meaning that the comment is unlikely to be toxic. A drawback of this model is that it does not specify the level of toxicity of a comment, it only gives the certainty with which the comment can be classified as toxic or not. Another drawback is that the algorithm is time-consuming due to its complex calculations and large data requirements. On our dataset, it took nearly 12 hours to compute. This makes this method hard to implement in a real- time moderation.

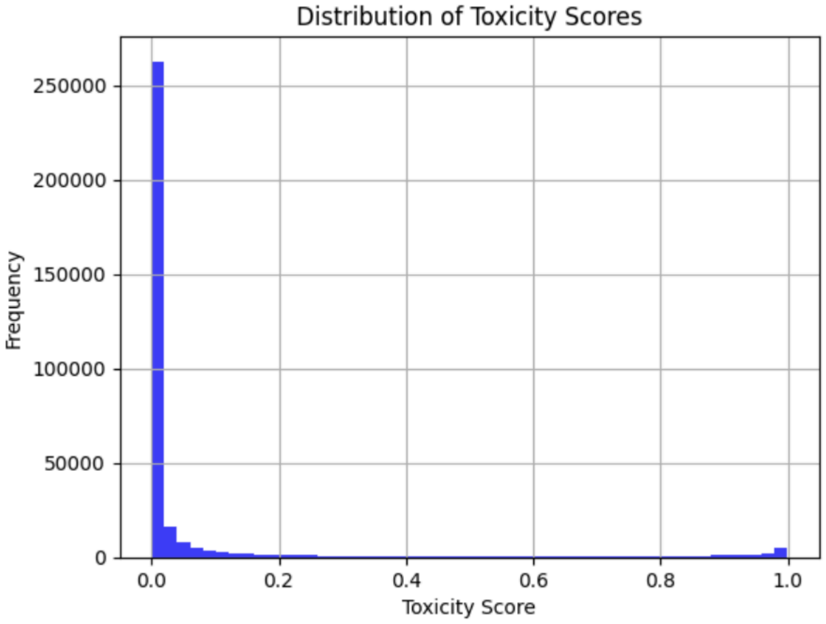


Figure 2 : Distribution of toxicity scores

The analysis of the toxicity score distribution reveals a predominant concentration of comments with low toxicity scores, indicating that most comments in the dataset are non-toxic. This distribution suggests that toxic comments are rare, which could allow moderators to focus their efforts more efficiently on these high-scoring comments. There is a clear need to add a toxicity threshold to filter comments.

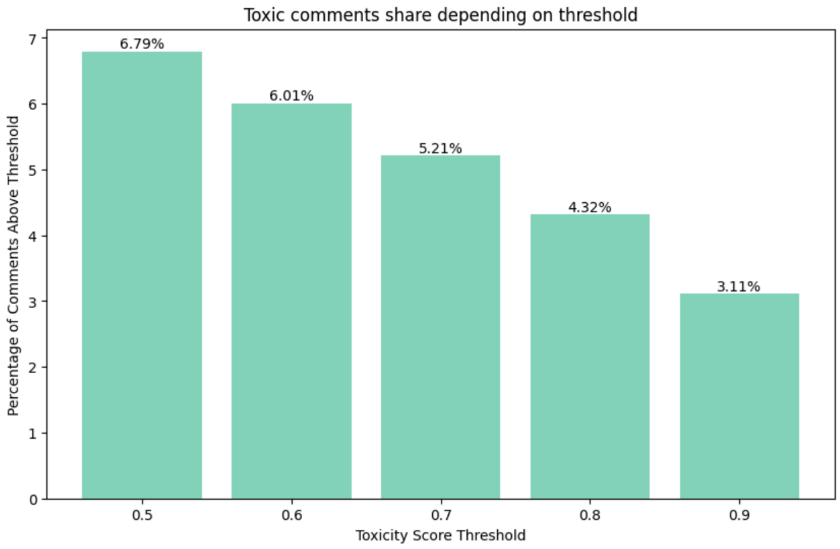
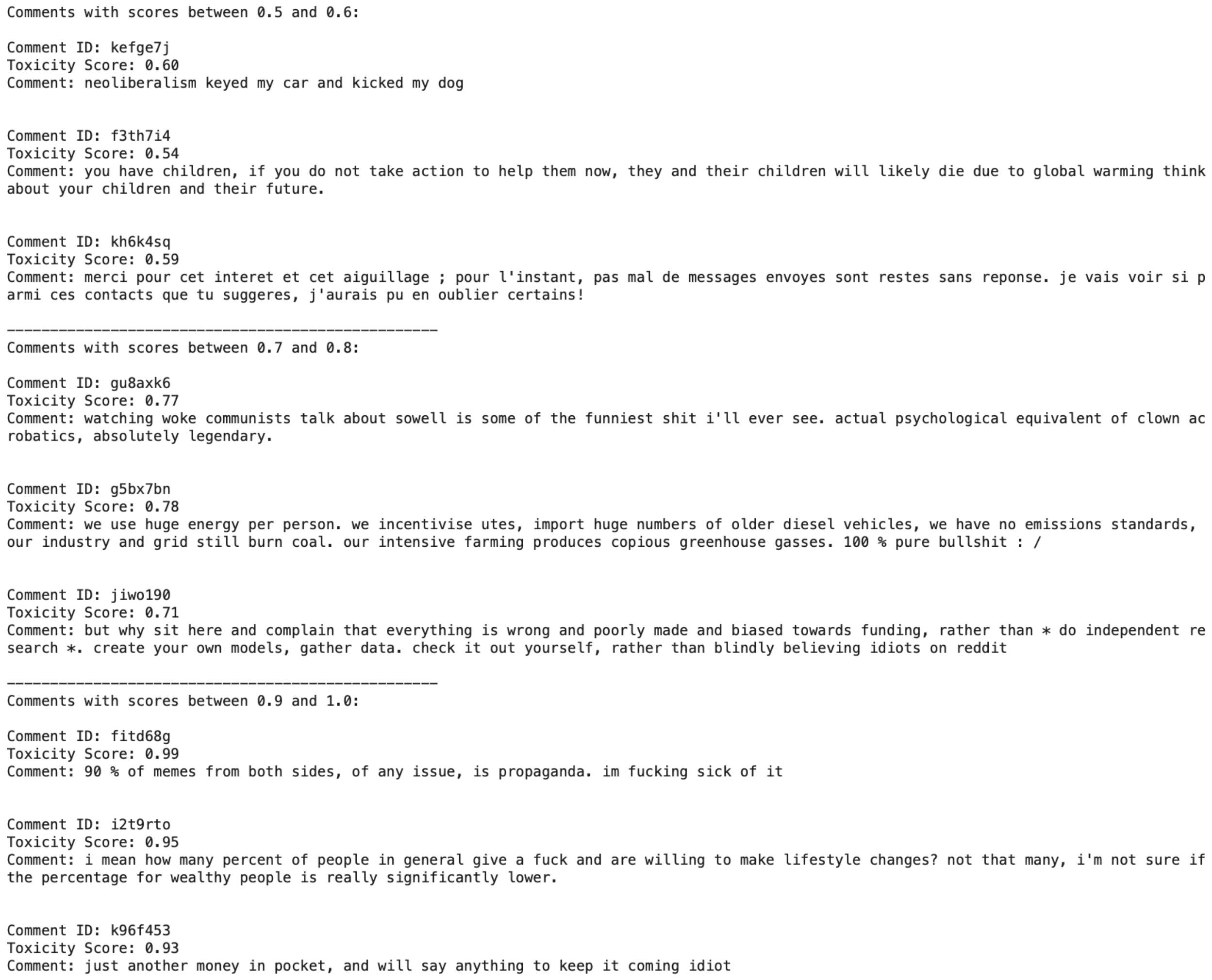


Figure 3 : Toxic comments share depending on threshold

The bar chart illustrates the percentage of comments classified as toxic across varying toxicity score thresholds, ranging from 0.5 to 0.9. As the threshold increases, there is a clear downward trend in the proportion of comments identified as toxic. Such information is critical for setting an appropriate toxicity threshold in moderation systems to balance between over-moderation and ensuring a safe online environment. The choice of threshold could significantly impact the volume of comments flagged for review, affecting both the workload of human moderators and the overall user experience on the platform.

As said before, our model does not quantify the level of toxicity of a comment, therefore a decision on deleting a comment or punishing a user has to be taken by a moderator. To illustrate what kind of message got which score, let’s print some comments falling between a certain score range :



```
Comments with scores between 0.5 and 0.6:

Comment ID: kefge7j
Toxicity Score: 0.60
Comment: neoliberalism keyed my car and kicked my dog


Comment ID: f3th7i4
Toxicity Score: 0.54
Comment: you have children, if you do not take action to help them now, they and their children will likely die due to global warming think
about your children and their future.


Comment ID: kh6k4sq
Toxicity Score: 0.59
Comment: merci pour cet interet et cet aiguillage ; pour l'instant, pas mal de messages envoyes sont restes sans reponse. je vais voir si p
armi ces contacts que tu suggeres, j'aurais pu en oublier certains!

--------------------------------------------------
Comments with scores between 0.7 and 0.8:

Comment ID: gu8axk6
Toxicity Score: 0.77
Comment: watching woke communists talk about sowell is some of the funniest shit i'll ever see. actual psychological equivalent of clown ac
robatics, absolutely legendary.


Comment ID: g5bx7bn
Toxicity Score: 0.78
Comment: we use huge energy per person. we incentivise utes, import huge numbers of older diesel vehicles, we have no emissions standards,
our industry and grid still burn coal. our intensive farming produces copious greenhouse gasses. 100 % pure bullshit : /


Comment ID: jiwo190
Toxicity Score: 0.71
Comment: but why sit here and complain that everything is wrong and poorly made and biased towards funding, rather than * do independent re
search *. create your own models, gather data. check it out yourself, rather than blindly believing idiots on reddit

--------------------------------------------------
Comments with scores between 0.9 and 1.0:

Comment ID: fitd68g
Toxicity Score: 0.99
Comment: 90 % of memes from both sides, of any issue, is propaganda. im fucking sick of it


Comment ID: i2t9rto
Toxicity Score: 0.95
Comment: i mean how many percent of people in general give a fuck and are willing to make lifestyle changes? not that many, i'm not sure if
the percentage for wealthy people is really significantly lower.


Comment ID: k96f453
Toxicity Score: 0.93
Comment: just another money in pocket, and will say anything to keep it coming idiot
```

Figure 4 : Printing some comments with their associated score

After reading some examples, we noticed that some comments with high scores are not that toxic to us and vice-versa. However, even though final decisions have to be taken by Reddit's moderator, we noticed that comments deserving to be moderated are more likely to have a score higher than 0.7.

With this threshold 5.21% of the comments, namely 17600 comments, should be reviewed by moderators. Those comments were made by 8483 different users.

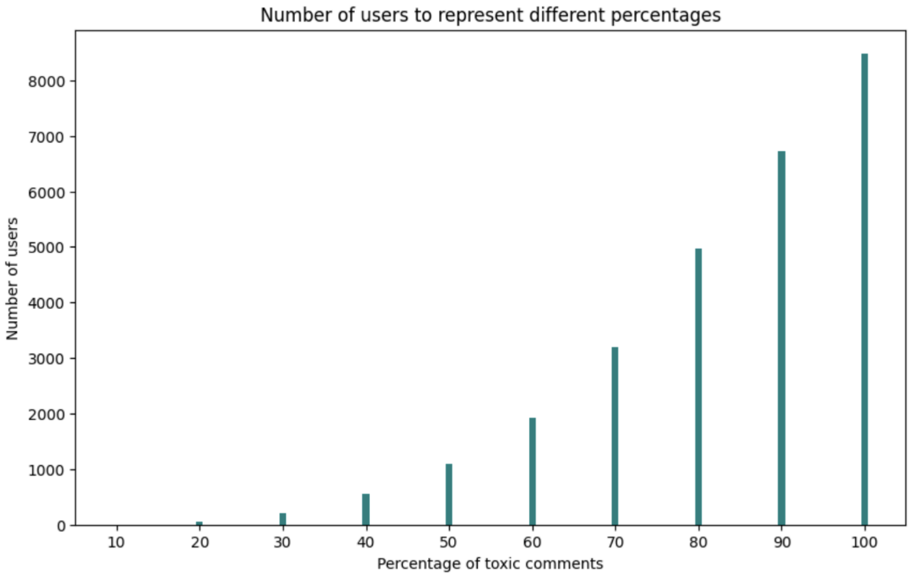
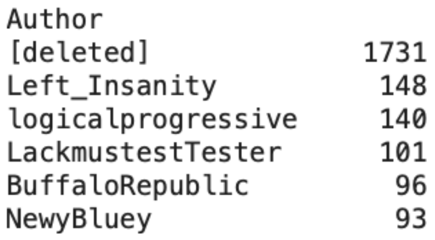


Figure 5 : Number of users to represent different percentages

However, some users write potentially toxic comments way more often than others. On this graph, we see that 50% of potentially toxic comments are made by only around 1000 users. This shows that moderation should be focused on some users. Just as an example, here is the list of the 6 most toxic users and the number of toxic comments they wrote, according to our model, one of them being all deleted accounts. This is a list that could be provided to the client.



```
Author
[deleted]               1731
Left_Insanity            148
logicalprogressive       140
LackmustestTester        101
BuffaloRepublic           96
NewyBluey                 93
```

Figure 6 : Most toxic users

This study was done on the Bert dataset, highlighting the toxicity of posts and comments. However, the Digital Service Act regulation puts forward many issues with content that moderators should remove: hate speech, violent content, disinformation, fake news, racist and sexist content, cyberbullying, … That is why the dataset used to train the model can change depending on the moderator's needs. for example, if a moderator wishes to detect fake news, we can do the same study using the LIAR dataset, or the ETHOS for hate speech. Plus, the dataset we have contains content on climate change, and it is possible to train a pre-existing model on climate change data, thus allowing an even more appropriate score on the content.

4. Detecting dangerous communities

People that participate in illegal activities tend to group together: in most cases, one cannot organize a terrorist attack alone, nor sell illegal products online alone. That is why, when considering illegal activity online, checking a user's interactions online is important. For that, we decided to create a weighted and undirected graph to represent all interactions between users in the dataset. The nodes of the graph are the different users, there is an edge between two users if they have already interacted together on Reddit (posts or comments), and those edges have an associated weight corresponding to the number of interactions between the two users. We considered that user interaction is attention worthy when at least 5 posts or comments were made between them, and as such the graphs do not take into account lower interactions.

Visualizing such a graph would not show anything of interest, but from this graph we can extract subgraphs corresponding to a user in particular, highlighting their interactions. Plus, we choose to keep interactions between the user's neighbors, and this can help detect communities or groups of users that share the same beliefs or participate in the same illegal activities. When moderating a user in particular, the moderator can then have access to potentially other users participating in illegal activities.

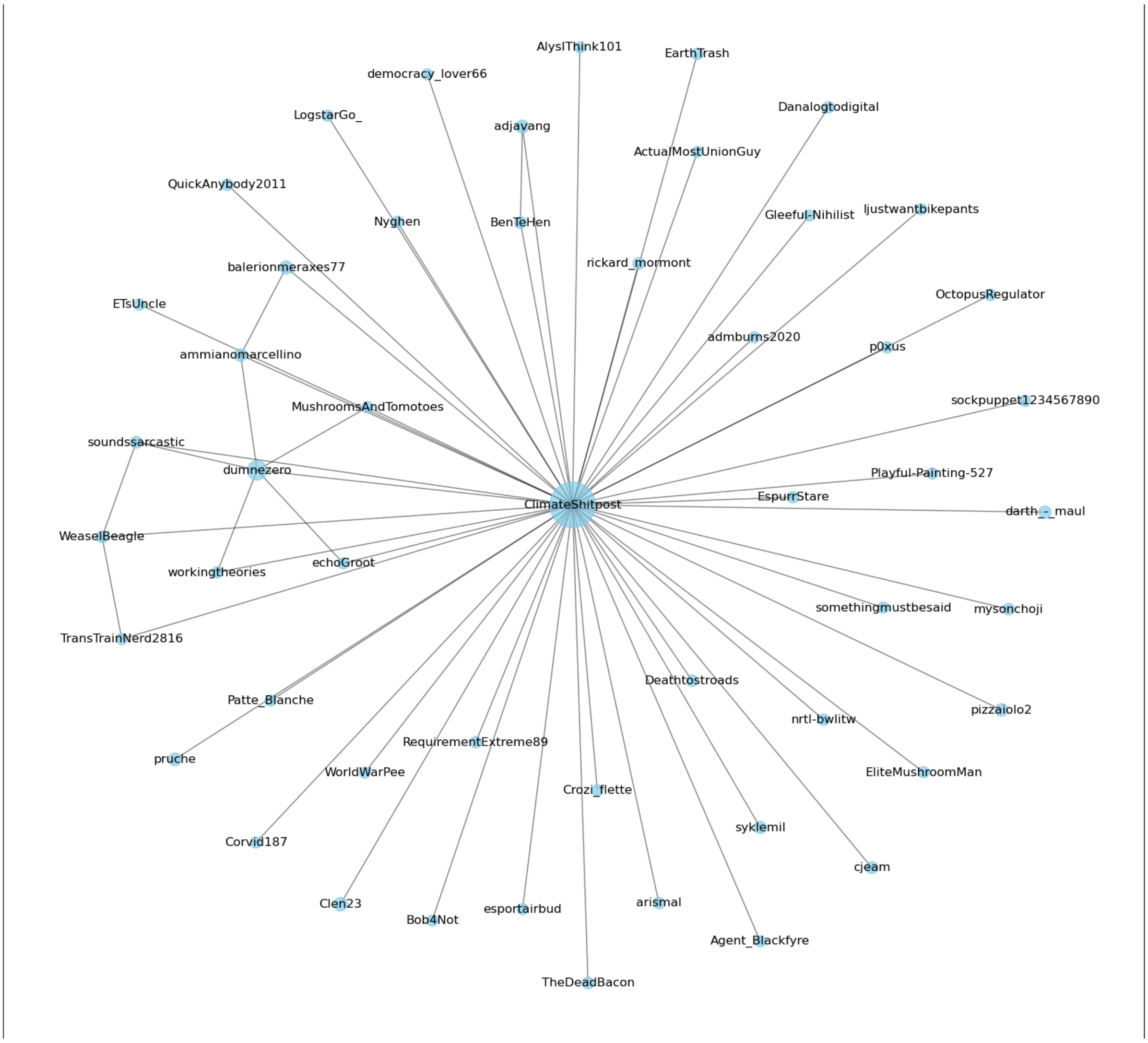


figure 7 : Subgraph of users ‘ClimateShitpost’ has interacted with

This can be visualized on the subgraph of the user ‘ClimateShitpost’. On this subgraph, we notice a small subset of users that interact not only with user ‘ClimateShitpost’ but together, and such groups should be noticed by moderators.

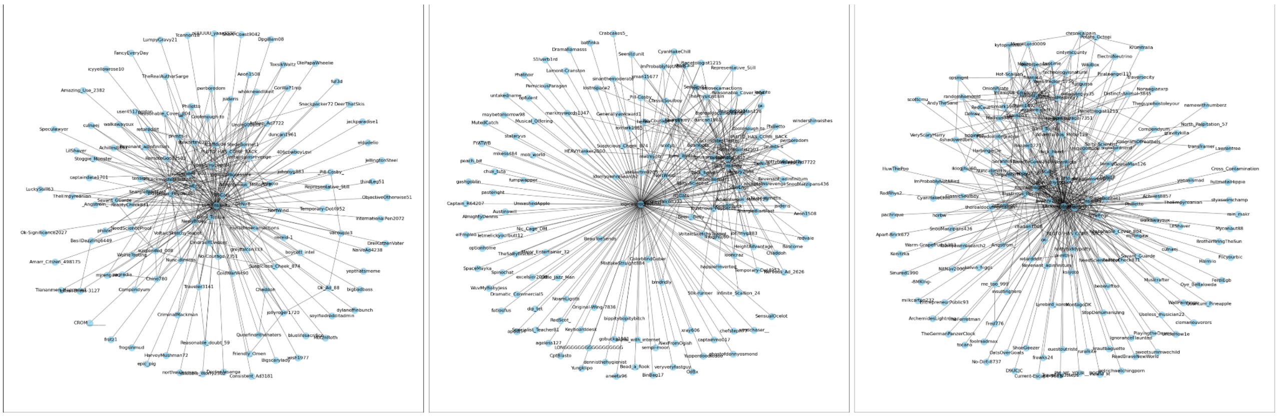


figure 8: Subgraphs of users ‘Left_Insanity’, ‘logicalprogressive’ and ‘LackmustestTester’, from left to right

The figure shows the subgraphs of the three most toxic users, according to our previously used Bert model. for the three users, we notice clusters that these users are part of. The existence of this cluster, along with other users part of it, is information that could be available to the moderator alongside a warning for one particular user or post/comment.

This study of clusters of users participating in illegal activities brings additional information to the moderator. Since the moderation work should be done as quickly as possible, before a post or comment has been seen by many people, it is hard to find users that will interact with such a post or comment, since it should be deleted beforehand. As such, having access to users one has previously interacted with can lead the moderator to remove other potentially dangerous users without having to keep illegal posts or comments online.

5. Conclusion

Using our model, we can give access to moderators on the toxicity score of posts and comments, as soon as they are posted, warning them on the ones that have the highest scores. Plus, depending on the moderator’s needs and the type of content (here, climate change posts and comments), our model can be adapted to detect other types of illegal content, such as the spread of fake news. The moderator can then decide if the content we have warned them about should be removed from the platform or not. This moderation work, although it can be done on previously posted content, can and should be put into place for content as soon as it was posted, to ensure only few users can see it. Plus, if a post or comment receives a quickly growing number of interactions, the moderator should also be warned (although our dataset does not give the timestamps of downvotes, for example). When warning a moderator about a post or comment, we can also warn them about the user that has posted it and their history: previous posts or comments with high toxicity scores. Plus, our graph study gives the clusters this user is part of, and the other users that can share the same illegal content online, thus allowing the moderator to remove many users at the same time and making their work easier and faster. As such, implementing our model for moderation on Reddit can help the platform follow the Digital Service Act regulation.