Report of analyzing the Impact of toxicity on Youtube Communities using a Toxicity Classifier

Professor Rafik Hamza
Tokyo International University
21224520 Jason Sebastian, 21229058 Ruchiraprakan Pamonpon
11-27-2024

Introduction

YouTube, as one of the largest platforms for content sharing and discussion, has become a medium for diverse opinions. Though, the amount of toxic comments, especially one that contains a huge amount of abusive language, hate speech, or personal attacks has increased by a lot, creating problems for both viewers and creators. In addition to encouraging negativity, these reduced the amount of positive participation.

Toxic comment classification could be a first step on improving the YouTube community. By detecting harmful content, platforms may improve moderation efforts, encourage better user interactions, and gain info on the tone of discussions around video contents. Ultimately, this can promote a more humane and better society by introducing positive energy. This report examines the development of a machine-learning classifier to detect toxic comments on YouTube, demonstrating its potential to improve content evaluation and community standards.

Model

The toxic comment classifier used RoBERTa model from Hugging Face's library that got improved on a dataset of label toxic comments from youtube videos, score output from this model ranging from 0 to 1 with 0 indicating no toxic content and likely having toxic content toward 1, testing those on Jigsaw dataset indicated good performance, getting an AUC-ROC score of 0.98 and F1 score of 0.76. These metrics indicate reliable detection capabilities among different states of toxic content. The model continuous scoring system indicated flexible range based on detailed needs.

Methodology

For our methodology, Mainly, we used Youtube's API to obtain the first 100 comments on youtube, then preprocess those data by changing emojis to closely matching texts to maintain meaning. Then analyze those texts by introducing a model from a hugging face that hasn't been trained to each comment and get a result in the form of a float number ranging from 0 to 1. A score closer to 1 displayed higher confidence in the toxicity element. By grouping scores to related comments we created our dataset. This methodology allows for an analysis of a larger dataset while still keeping the basic of classification through a continuous system rather than binary labels

Datasets

Dataset 1 is from a Youtube video uploaded by a Youtuber named Atrioc that showed his reaction throughout the election day. Atrioc is a Business and Gaming Youtuber that has publicly shown his interest in Kamala Harris winning but has mentioned that he thinks Donald Trump will win. This Youtuber was chosen as his community is politically diverse and is opposite to the next Youtuber's community.

Dataset 2 is from a Youtube video that was uploaded by a Youtuber named Hasanabi on his reaction during election day. Hasanabi is a political Youtuber that showed interest in Kamala Harris winning and believes that Kamala Harris will win on election day. The community has the same political beliefs as the Youtuber and is more likely to contrast in toxic comments than the Atrioc's community.

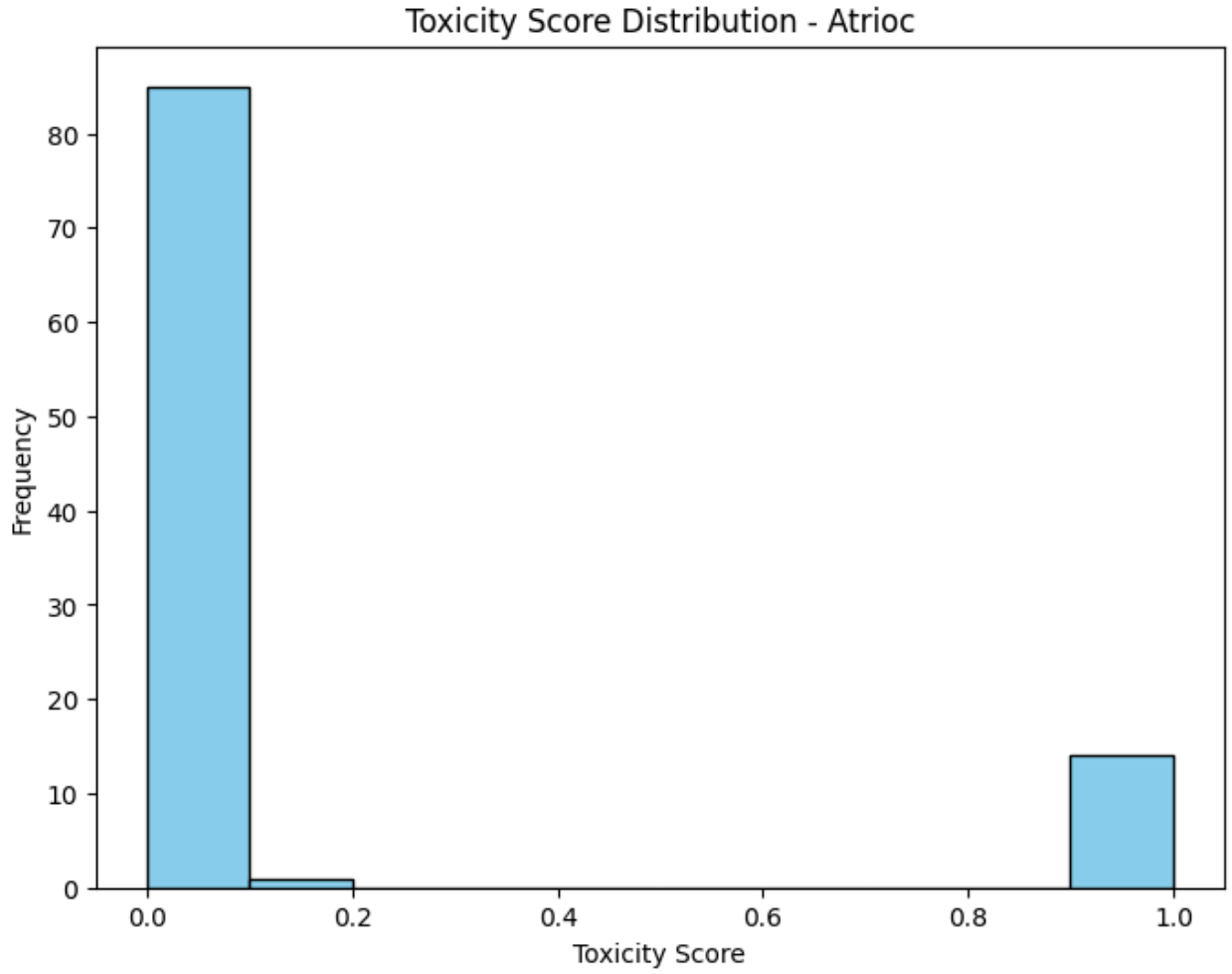
Dataset 3 is from a Youtube video of Kylie Jenner's last video before she stopped uploading on Youtube. The video is a Behind The Scenes video that does not align with any political views and can be considered a neutral video. A neutral video, in this definition, would be a video that does not promote hate or push an agenda but rather just a video purely for entertainment. Kylie Jenner was selected as she is a celebrity that recently quit from uploading Youtube videos as she was getting a substantial amount of toxic comments from individuals that disliked her.

Dataset 4 used comments from Dr Disrespect's latest youtube video with comments enabled. He is known as a problem Youtuber as he was caught cheating on his wife in the past

along with several allegations of grooming. These controversies led to multiple breaks from streaming, with recent attempts to return to streaming and uploading. The data has comments that require context to recognize as toxic. This video was chosen to highlight how difficult it will be for a model to predict toxicity without the historical context of the Youtuber.

Observations

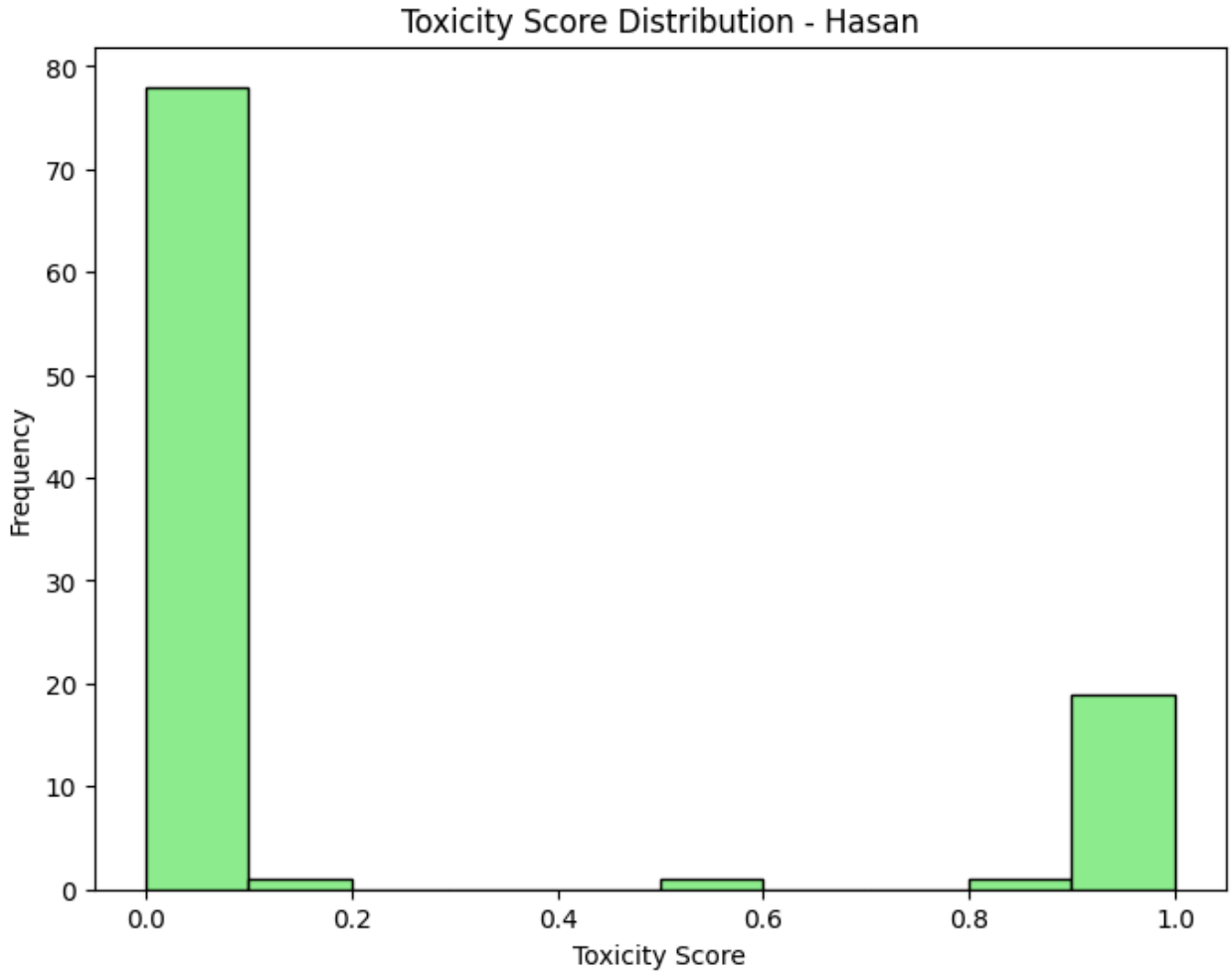
Atrioc



The video is about a Business Youtuber, who is pro-Kamala but believes Trump will win, reacting to the US Election results. In the first 100 comments, it is shown that **14 of them were likely to be toxic** with a **Max toxicity of 0.999586**. **The average toxicity of all the comments was 0.143** which was the second lowest in the dataset. The histogram shows that the comments

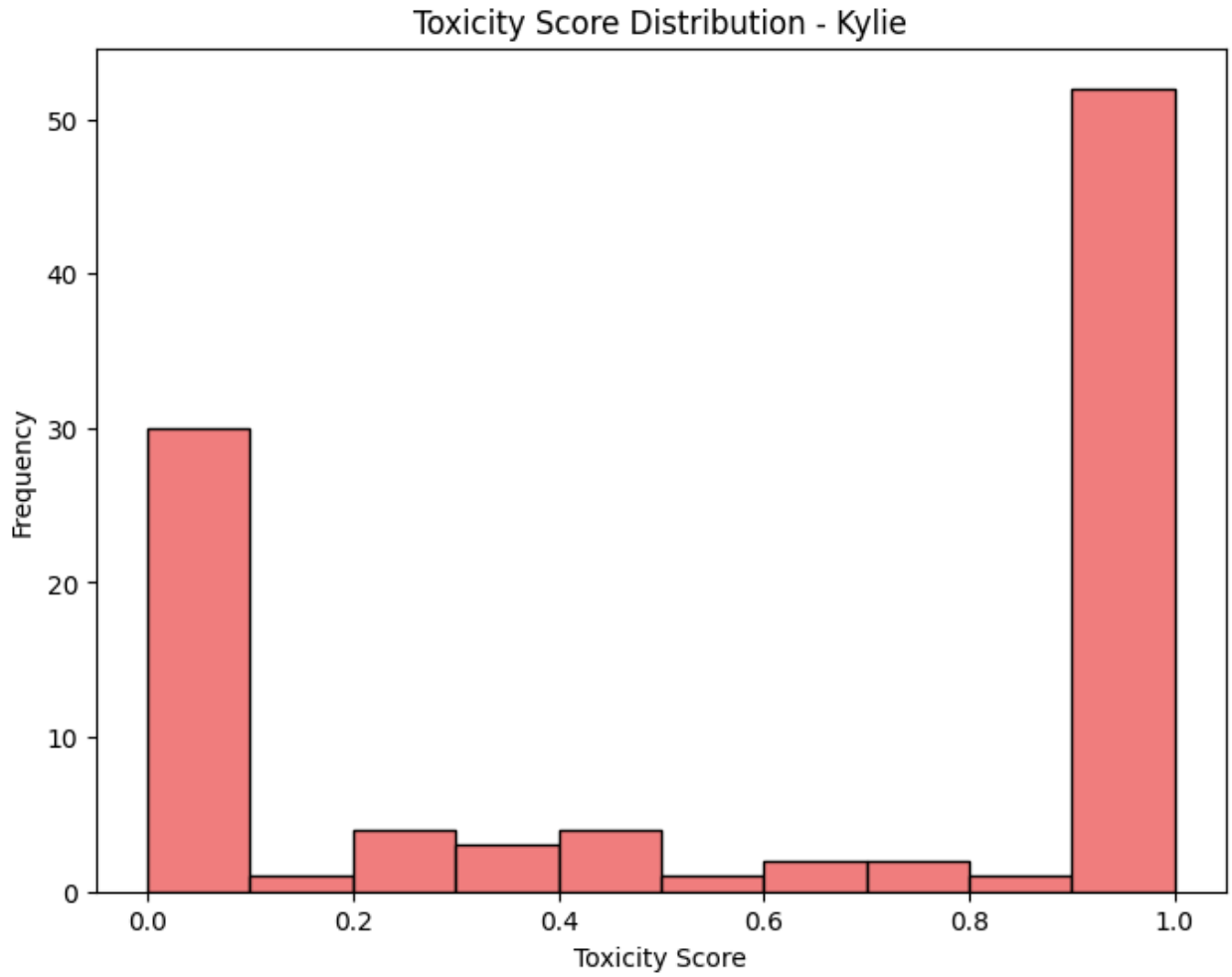
can be predicted as toxic and non-toxic, suggesting the model was confident in predicting the comments to be likely toxic or non-toxic.

Hasan



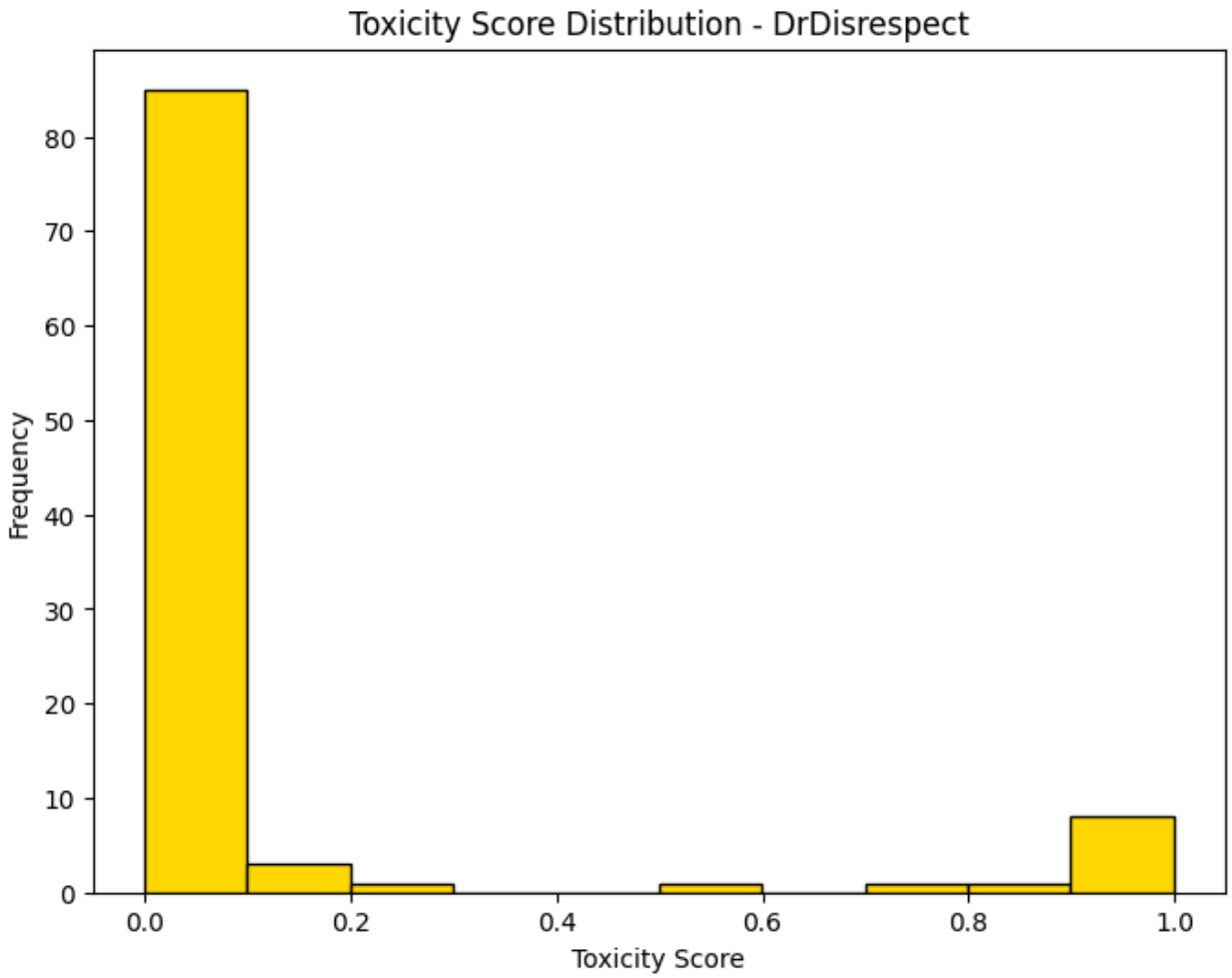
The video used is about a political Youtuber, who is also pro-Kamala and believes Kamala will win, reacting to the US Election results. In the first 100 comments, it is shown that **21 were likely to be toxic** with a **Max toxicity of 0.999597**. **The average toxicity of all the comments was 0.206**. The histogram showed similar to Atrioc's with a few comments being predicted close to 0.6 which suggest the comment was predicted to be only mildly toxic.

Kylie



This video was the last video uploaded by Kylie Jenner, a celebrity, which was a behind the scenes video of an event she was in. In the first 100 comments, it is shown that **58 of them were likely to be toxic** with a **Max toxicity of 0.999597**. **The average toxicity score of the comments was 0.60**, being the highest. Kylie's histogram also showed the most spread out toxicity score between the comments which can suggest that there was more trolling. Trolling is defined as an act of online harassment to upset or annoy others. It can suggest trolling because the toxicity scores are around 0.5 which may indicate that the comment is mildly toxic or infuriating but not harmful or abusive.

Dr Disrespect



This video was chosen as data that required context to understand the toxicity in the comments. Dr Disrespect is a Youtuber that has taken a break from video creation multiple times as multiple allegations have been reported of him. The video comments would be difficult to detect toxicity as, once again, requires context to understand. With **11 of the comments likely to be toxic**, with a **Max of 0.998969** and an **Average of 0.108** it has the lowest values of all features. The histogram shows that there are a lot of comments the bot has predicted to be non-toxic but in this case, it can suggest that without the historical context, it is closer to uncertainty.

Analysis

Streamer	Number of Toxic Comments	Max Toxicity Score	Average Toxicity Score
Atrioc	16	0.999219	0.164556
Hasan	21	0.999597	0.206275
Kylie	59	0.999638	0.610181
DrDisrespect	11	0.998969	0.108397

The data shows that toxicity is present in all types of Youtuber communities and videos. The minimum being 11 toxic comments out of 100 comments used shows that at least 10% of comments in videos are likely to be toxic. The target of the toxicity can change the effect of the toxicity towards the Youtuber's communities.

The videos used to represent Hasan and Atrioc were both videos of the election but the differences were what type of communities they had. Hasan is a political Youtuber with a community that has the same opinion as him and Atrioc is a business Youtuber who does not push for any political agenda. After Kamala Harris' defeat, it is evident that Hasan's video had more toxic comments as well as a higher maximum toxicity score. This suggests that Hasan's community, having a shared political belief, may have contributed to the higher number of toxicity when Donald Trump won. In contrast to Hasan's fans, Atrioc's fans were more business oriented which makes it likely that the toxicity is the result from fighting between fans of different political opinions. This shows opposite targets of toxicity. One is toxicity towards another community, while the other is toxicity within the community.

Kylie Jenner and Dr Disrespect are both targets of toxicity from individuals of different groups towards Youtubers. Kylie Jenner's video was a behind the scenes video during an event she was present at. The video was not meant to incite or push for any agenda but rather just a video to pull views. The video received 59 toxic comments with an overall average of 0.61 toxicity score from all the 100 comments, having the highest average toxicity score from the 4 datasets. These are likely not from her community but individuals commenting on her video. The

effect of the toxicity from this video likely caused her to stop uploading Youtube as this video was her last video before quitting Youtube altogether.

Dr Disrespect is a Youtuber that has been receiving hate after multiple allegations surfaced and after he announced that he cheated on his wife. He attempts to return after every incident and receives many hate comments. Since the model cannot receive the context of the comments, it can only detect the comments that are directly insulting him with no mention of past events or allegations. For the model to identify the toxic comments, it will require the historical context and the sarcasm from the comment. This dataset was picked to show the restrictions of the model when evaluating comments that require context to understand the toxicity. Because of this, the model could only accept 11 comments as toxic with an average of 0.108.

```
Neutral_comments 89
Moving to Rumble has to be the biggest slap in the face to all the people who became members on your Youtube... Didn't even g
ive the people a heads up.
great now he is streaming on rumble and not youtube anymore....
Doc you ought to double down. Those virgins don’t understand real life struggles. There’s pleeeeeeeenty worse. You only had a
slip up.
STAY STRONG
Why is he not banned
Dr445
DOC GIMMIE THJAT 25K!!!
Opens Video to watch Him Play::cross_mark:
Opens Video to read the Comments::check_mark_button:
Dr DisRespect. No you can't have my sister but you can have some :cupcake::cupcake::cupcake::cupcake::cupcake::cupcake: yummy
yummy
turned off comments WONT stop us
12:06  bro has lots of practice in creating lullabies for babies
WAKE UP @DrDisRespect we are all ready to see you in a hospital gown with your vest overtop laying in bed playing :face_with_
tears_of_joy:
Dr Disrespect is too busy talking to 6 year old girls
Hey doc, whens the next video coming out? These ones are getting pretty old and its not like you to wait for things to get ol
d...
isnt bro supposed to be in prison :loudly_crying_face::folded_hands:
Certified lover boy :clown_face:
Chris Hansen will get you
Ngl Doc im glad youre back youre gameplay is relatable and you made game for me fun im glad i found you
How tf do ppl watch this guy? 2:15
12:18 Dr. Disrespect walking into a Toys R Us be like
Uhhhh 2:37
we still disrespecting you
He’s cooked
Aren’t you supposed to be in jail ?
Shoutout to DrDisRespect. It would be awesome to SOCOM remastered and brought back. The beginning of console online gaming fo
r PS2. Hahaha, to my knowledge at least
1:57 didn’t you cheat on your wife with some men?
Dr respect is better than
Creep
```

Revealing the neutral comments can also show that there are comments that require historical context. The comments about sisters and little girls are mentioning his accusations of inappropriate conduct involving minors in the past (Murray, 2024). The cupcake emojis also require context of another accusation on a different individual.

Conclusion

In conclusion, toxicity appears to be pervasive in YouTube's current landscape, yet its impact can be either negative or positive, depending on the target of the toxicity. Positively, toxicity can be a source of information to allow other individuals to know about a community before joining them. But it can also negatively affect a Youtuber's community. If connected to the correct program, it would be able to automatically delete toxic comments before it reaches the Youtuber. In the case of Kylie Jenner, she likely would have continued posting videos instead of getting affected by the toxicity.

Future Considerations

Our main goals of future research should mostly focus on improvements of toxicity classification models including but not limited to, accuracy or other adaptive traits. This can help reduce bias and ensure that this can work well across diversity of language, group or culture. The process of improving a model's accuracy can be broken down to the way it takes different phrases, idioms and changing how to interpret those data, refining and adapting continuously to be in trend with recent wording used by humanity. In which users can also help assist different inputs and further train the model.

References

Murray, C. (2024, August 27). *Dr. disrespect controversy: YouTube reportedly suspends monetization after streamer admits sending "inappropriate" messages to minor*. Forbes. https://www.forbes.com/sites/conormurray/2024/06/28/dr-disrespect-controversy-youtube-reportedly-suspends-monetization-after-streamer-admits-sending-inappropriate-messages-to-minor/