YouTube Gaming Comment Toxicity: A Comparative Analysis of Action and Non-Action Games Video Comments

Chesie Yu,[1] **Hongfan Lu,** [1] **Bella Wei** [1]

[1] University of Washington
cyu909@uw.edu, hongfl@uw.edu, yuewei6@uw.edu

Abstract

Toxicity in the gaming community is a prevalent issue within the gaming industry, posing serious concerns for individuals' mental well-being and societal stability. While existing studies on gaming toxicity primarily focus on interactions within in-game platforms, our study approaches it from the perspective of game observers. We chose YouTube as our targeted platform and utilized Google Perspective API and NLP sentiment analysis to investigate two main aspects of toxicity within the comment section:

1. The impact of game genres (action and non-action).
2. The influence of video transcripts on comment sections.

Our findings reveal that:

1. Action games tend to elicit more toxic comments.
2. Toxic video transcripts are more likely to generate toxic comments.

However, we also discovered that such correlations are not as strong. We speculate that there are more factors influencing toxicity in the YouTube comment section. In future, we will conduct more comprehensive study and we hope our findings provide practical insights to video creators, video managers, and the gaming industry to enhance the harmony of gaming communication.

I. Introduction

Motivation

Gaming is a widespread interest in modern society, particularly among children and teenagers. In our busy everyday lives, playing games serves as a significant means for people to relieve stress and find joy. As a result, the gaming industry and the corresponding video sector are thriving. However, the issue of toxicity stemming from gaming has become increasingly serious. This toxicity takes various forms, including profanity, discrimination, and sexual harassment, and can lead to significant consequences such as cyberbullying, mental health issues, and even antisocial behavior. In particular, this toxicity can spread rapidly on social media platforms. We are motivated to study the factors contributing to this toxicity and, by addressing them early on, contribute to the long-term development of gaming platforms.

Research Objective

This study aims to address two research questions within the YouTube gaming community:

- Does the genre of games (action and non-action) significantly impact the toxicity in the comment section?
- Does the content of the videos have a substantial influence on the toxicity in the comment section?

Through this research, we hope to generate valuable insights for game development, social media promotion, and platform management.

II. Data

Data Collection

Using YouTube as our primary data source, we compiled a dataset containing 136,463 comments from 1,407 unique videos across 32 channels. Our data collection strategy follows a three-step process.

Keyword Selection To differentiate between the target game genres, we devised two sets of keywords representing popular games within action and non-action categories. The keyword sets are outlined as follows:

- **Action Games:** {*call of duty*, *gta*, *the last of us*, *god of war*, *red dead redemption*, *assassin's creed*, *star wars jedi*, *resident evil*, *cyberpunk*, *fallout*, *tomb raider*, *elden ring*}
- **Non-Action Games:** {*minecraft*, *pokemon go*, *just dance*, *it takes two*, *uncharted*, *brawl stars*}

Channel Selection From Social Book's Top 100 Gaming YouTubers, we curated a list of 32 channels that predominantly produce gaming content in English through a manual validation process. We extracted the channel IDs and tagged them with binary labels "english" and "gamer", ensuring their relevance to our research focus on the English-speaking gaming community.

Data Collection We then input the filtered channel IDs into our data collection pipeline. Utilizing the YouTube Data API and yt-dlp, we obtained 30 videos per category from every channel, through pre-defined keywords for action and non-action games. Afterward, we collected the 100 most relevant top-level comments for each video. This procedure yielded 26 features associated with channels, videos, and comments, capturing metadata, statistics, and text-based content including subtitles and comment text.

Data Preprocessing

Preprocessing was subsequently performed to maintain the integrity and relevance of our collected data.

Data Cleaning Initial data cleaning steps involved handling missing values, detecting duplicate and invalid entries, and standardizing data formats. We opted for deletion to address the small fraction of missing entries (0.0585%) to ensure the dataset's completeness for analysis.

Feature Engineering By transforming existing variables, we derived several new features relevant to our analysis. This included identifying the games referenced in video titles and measuring the proportion of censored words in subtitles through regular expressions. Doing so enables game-specific analysis and offers a supplementary dimension to video toxicity assessment.

Toxicity Annotation

Acquiring the toxicity labels is imperative to our analysis, yet the manual annotation of approximately 130,000 comments is infeasible due to the large scale and limited resources. Consequently, we leveraged the Perspective API for a proxy to true labels, quantifying the level of toxicity in videos and comments. The API evaluates the raw text and assigns a perceived toxicity score from 0 to 1 across 6 subtypes, with higher scores indicating a greater likelihood of toxic content.

We faced two major challenges during this phase:

Quota Limit The Perspective API enforces a quota limit of one query per second (QPS) for each project, presenting a significant limitation for our data processing needs. To enhance our query capacity, we developed a throttling management strategy incorporating key rotation and exponential backoff, cycling through 10 keys along with their respective pre-configured clients. The exponential backoff mechanism introduces a retry logic following any server errors, incrementally extending the delay between subsequent requests and minimizing the likelihood of successive failures. These measures collectively reduce the projected processing timeframe from roughly 3.37 days to under 6 hours.

Byte Limit The Perspective API is optimized for shorter, comment-length texts similar to its training corpus. It demonstrated limitations when tasked with analyzing video transcripts larger than 20,480 bytes. To navigate these limitations, we segmented the transcripts into overlapping chunks of 100 words, preserving the context and accounting for the fact that these chunks are not independent. Subsequently, we evaluated the weight of each chunk and assigned lower weights to overlapping words. This allowed us to calculate a weighted average of toxicity scores for the segments as the ”perceived video toxicity”.

Text Preprocessing

We employed NLP-powered text preprocessing methods to refine our corpus for analysis. Using Perspective API's detectedLanguage attribute, we first excluded non-English comments to align our study with the English-speaking gaming community. We then incorporated a two-tier text cleaning approach, tailoring preprocessing needs for different sentiment analysis tools.

Contraction Expansion and Noise Removal The initial stage focuses on contraction expansion and noise elimination. Using the contractions library, we expanded contractions to their full forms, enhancing text clarity for analytical tools. Additionally, URLs, mentions, hashtags, and new-line characters were removed to reduce noise in the data.

Normalization and Punctuation & Stopword Removal In this step, text normalization is carried out by converting all texts to lowercase for uniformity. Furthermore, non-alphabetic characters and common English stopwords that lacked significant information were removed to focus on relevant information.

Sentiment Analysis

For a multifaceted exploration of the emotional dynamics within YouTube comments, we computed the sentiment scores using VADER, TextBlob, and Empath. Specialized for social media text analysis, VADER interprets nuances such as punctuation, slang, and emojis. Implemented on minimally processed subtitles and comments, VADER assigned a compound score ranging from -1.0 (most negative) to 1.0 (most positive), reflecting the overall sentiment for each document.

TextBlob and Empath analyses were performed on more extensively cleaned texts. The polarity score from TextBlob measures sentiment on a scale from -1.0 to 1.0, providing a quantitative assessment of emotions. For Empath analysis, we collected all 194 categories, aiming for a more nuanced understanding of the prevalent themes within YouTube gaming comments.

Our final corpus consists of 124,704 comments and 1,307 videos, characterized by 448 features. The following section will discuss the analysis and key findings from our study.

III. Analysis and Findings

Our analyses were achieved through 3 phases, each was building on top of each other. In the section below, we provide a brief overview of the steps we took and our observations made.

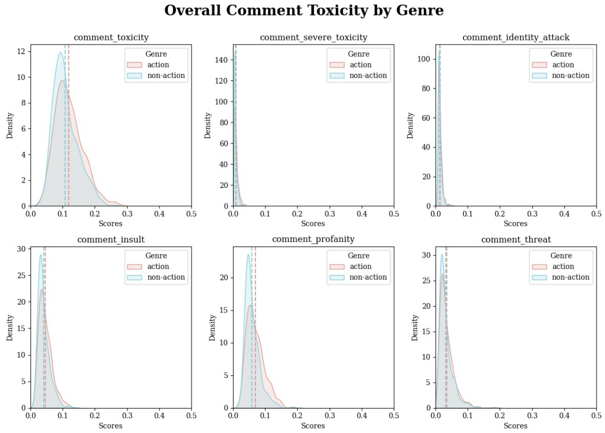
Phase 1

There were a few caveats: we use a predefined keywords list to fetch videos, which may not be representative enough; moreover, we retrieve 100 comments under each video, provided by the default order on YouTube, which may introduce bias.

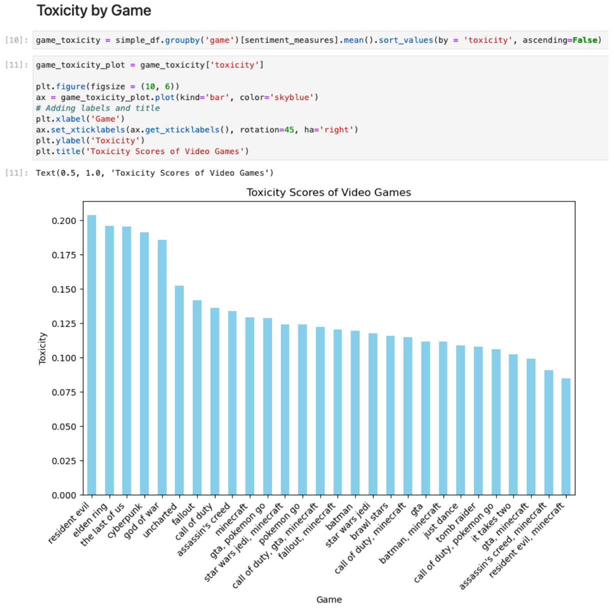
Our phase 1 analysis was conducted with the following two research questions in mind:

- Do videos of action games arouse significantly more toxic comments than non-action games on YouTube?
- Moreover, the occurrence of profanity might be higher in the action based gaming videos.

To answer question one, we first plotted the distribution of Toxic comments among the six available dimensions. The cutoff point we chose was toxicity bigger than 0.3. Among all dimensions, the label 'toxic' was the most often observed one; we then move on to plot the sentiment (VADER, TextBlob and Empath) density plots but we saw that the overall distribution of action and non-action video almost overlapped. However, the mean of either action or non-action comment sections' sentiments differs slightly, with action game video comment sections having slightly higher Vader neg values.



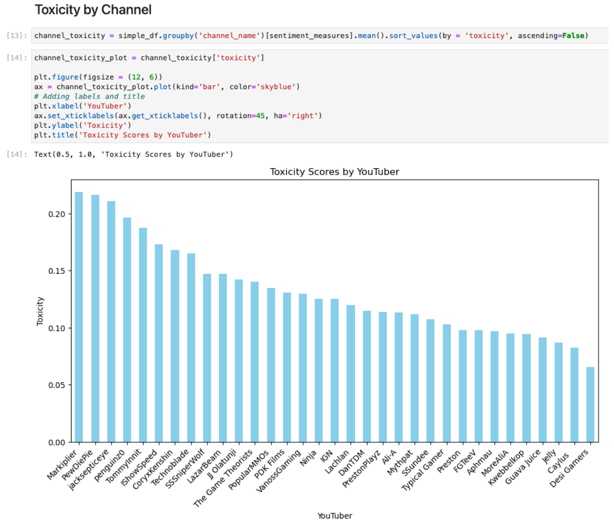
To provide more clarity, we then group the whole dataset of all comments by Game. When we plotted out the bar chart and ranked the bar in a descending order of toxicity scores provided by Perspective API, we indeed observe action game comment sections tend to have higher toxicity scores and thus take the left part of the bar graphs. While non-action games like Minecraft, tend to squeeze on the lower Toxicity scores side.



Toxicity by Game

```
[10]: game_toxicity = simple_df.groupby('game')[sentiment_measures].mean().sort_values(by = 'toxicity', ascending=False)

[11]: game_toxicity_plot = game_toxicity['toxicity']

      plt.figure(figsize = (10, 6))
      ax = game_toxicity_plot.plot(kind='bar', color='skyblue')
      # Adding labels and title
      plt.xlabel('Game')
      ax.set_xticklabels(ax.get_xticklabels(), rotation=45, ha='right')
      plt.ylabel('Toxicity')
      plt.title('Toxicity Scores of Video Games')

[11]: Text(0.5, 1.0, 'Toxicity Scores of Video Games')
```

After analyzing by game, we then move on to group all rows by channel (or YouTuber's comment sections) to investigate any potential nuances that could be brought in by the individual characteristics of the presenters. Our bar plot shows the most toxic YouTuber labeled in our analysis 1 "Markiplier" has a toxicity score of almost three times that of "Desi Game". To investigate the individual differences, we then plotted the proportion of Action videos in the overall posting ranking by YouTube channels. However, we were unable to find the pattern that higher proportion of action videos in one channel bring a higher toxicity level in the comment sections.



Toxicity by Channel

```
[13]: channel_toxicity = simple_df.groupby('channel_name')[sentiment_measures].mean().sort_values(by = 'toxicity', ascending=False)

[14]: channel_toxicity_plot = channel_toxicity['toxicity']

      plt.figure(figsize = (12, 6))
      ax = channel_toxicity_plot.plot(kind='bar', color='skyblue')
      # Adding labels and title
      plt.xlabel('YouTuber')
      ax.set_xticklabels(ax.get_xticklabels(), rotation=45, ha='right')
      plt.ylabel('Toxicity')
      plt.title('Toxicity Scores by YouTuber')

[14]: Text(0.5, 1.0, 'Toxicity Scores by YouTuber')
```

This finding challenges the common assumption that game category is a strong factor which influences toxicity levels and sentiment. It prompts a deeper exploration into the intricate factors contributing to online gaming toxicity. One possible reason is the gaming observer is a different group from game players. Another reason is the game culture in the YouTube gaming community is overall positive and kind, which makes the comments fairly positive.

Our observations from phase 1 motivate us to obtain the transcript level information, which represents the content, language habit, YouTuber characteristic and topic. We would like to investigate the relationship between transcript toxicity and comment section toxicity.

Phase 2

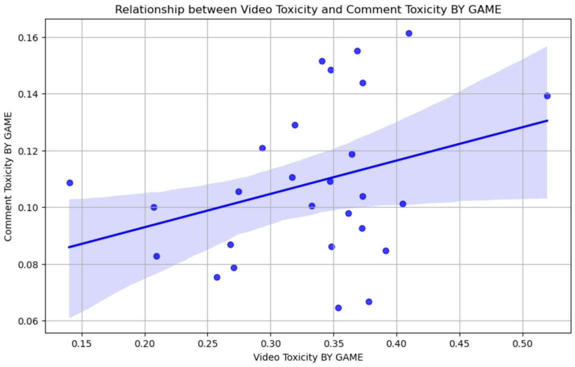
In this phase, we focus on investigating the relationship between video content and the comment section. Our hypotheses are:

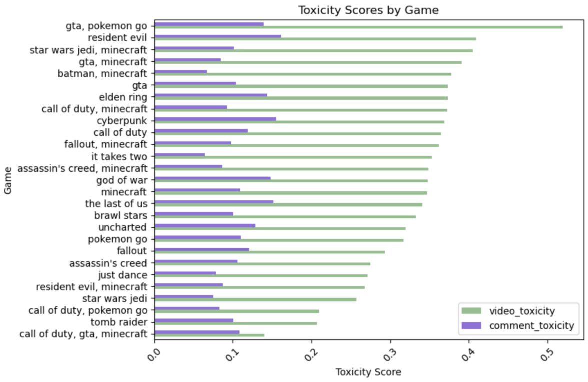
- Videos with higher toxicity in the transcript are likely to elicit more toxic comments.
- YouTubers who use more toxic language are likely to trigger more toxic comments.
- Topics in toxic videos are expected to involve more violence, war, and fights compared to harmonious videos.

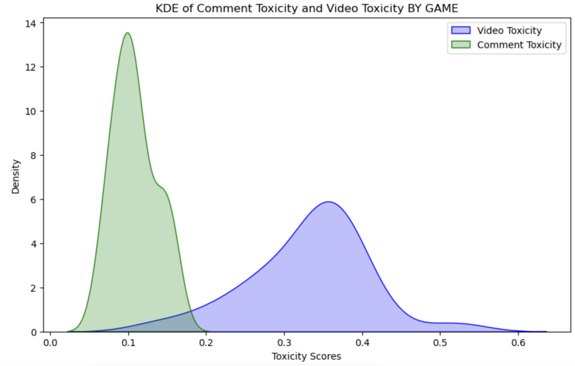
We updated our data collection methods and included the finer details on the Empath topics (detailed above in the Data Collection section).

After adjusting the threshold from 0.3 to 0.5 for labeling a comment as toxic, the overall numbers of toxic comments decreased from around 20000 to 2739 (around 2.2% of all); around 5144 (4.1%) numbers of video transcripts are marked as toxic. Both the KS test and Levene test statistics shows that the two distributions, comments and transcripts, are having different distributions and variances.

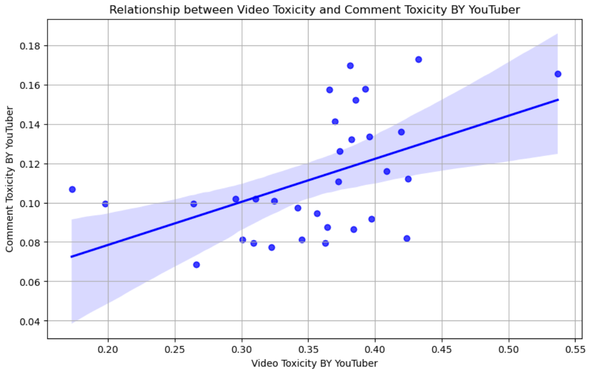
Like in phase 1, we group the dataset by game and run a regplot to the aggregated data. Our graph showed that there is a positive correlation between the video toxicity and comment toxicity. Although the scatters are sparse (compared to the graph grouped below by channel/youtuber), regplot was able to identify a positive relationship. This tells us that not only transcripts and comments toxicity levels are related, but also that some games are indeed more toxic than the others, although we have yet to test this hypothesis statistically. This fact is observed, INDEPENDENT of YouTuber's individual effect on comment sections.

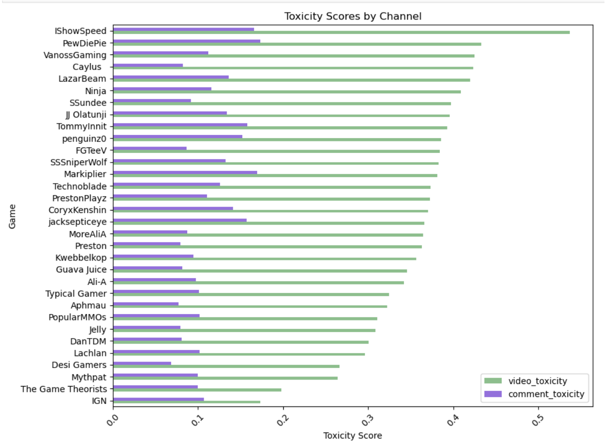


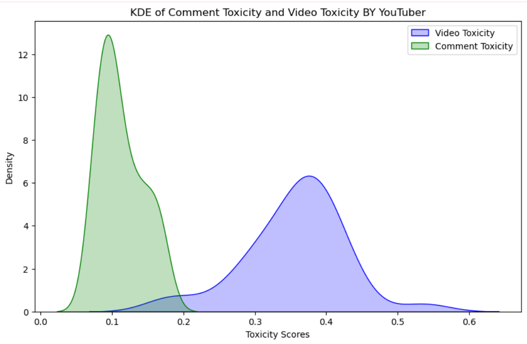




We then group the dataset by channel, and the graph is a much denser looking scatter and regplot and the positive slope of the regression line is steeper compared to the scatter plot for the game. This tells us that each youtuber's expressions, word uses and his or her presence are likely to be more explanatory for the toxicity correlations.



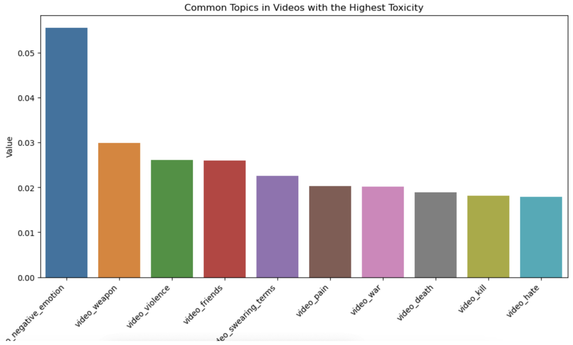


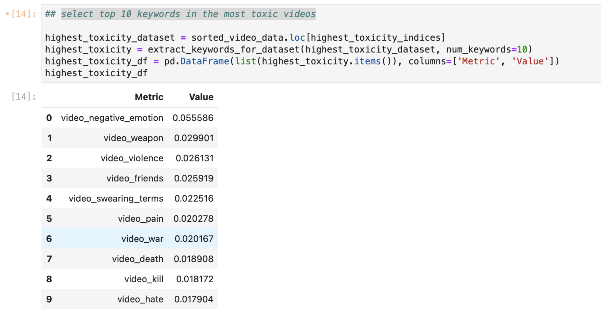


There are a few observations from two plots above:

1. The scatter plot is showing a much denser pattern between transcripts and comment section when we group by channel:
2. This tells us that each youtuber's expressions, word uses and his or her presence are likely to be more explanatory for the toxicity correlations.

Next, we would like to investigate Empath keywords most commonly used in toxic or non-toxic videos. Surprisingly, the common topics in the most toxic and least toxic videos exhibit clear differences and patterns. Toxic videos frequently cover negative topics such as ’violence,’ ’war,’ and ’kill,’ while non-toxic videos predominantly feature positive topics like ’music,’ ’dance,’ ’art,’ and ’fun.’ Intuitively, it can be observed that the topics in toxic videos are more negative in nature compared to non-toxic videos.

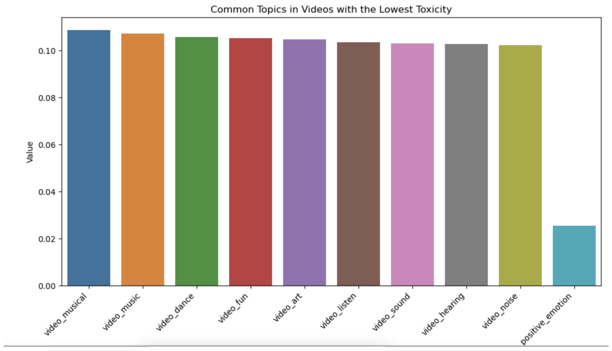


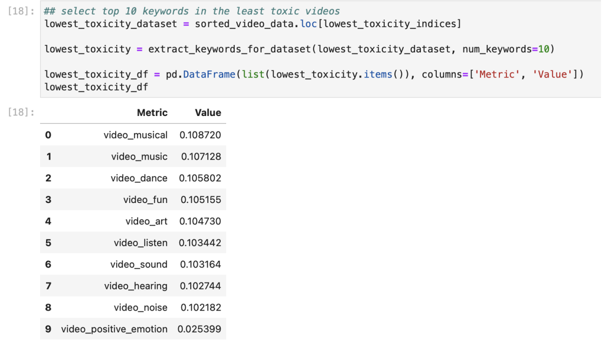


•[14]:
```
## select top 10 keywords in the most toxic videos

highest_toxicity_dataset = sorted_video_data.loc[highest_toxicity_indices]
highest_toxicity = extract_keywords_for_dataset(highest_toxicity_dataset, num_keywords=10)
highest_toxicity_df = pd.DataFrame(list(highest_toxicity.items()), columns=['Metric', 'Value'])
highest_toxicity_df
```

[14]:

	Metric	Value
0	video_negative_emotion	0.055586
1	video_weapon	0.029901
2	video_violence	0.026131
3	video_friends	0.025919
4	video_swearing_terms	0.022516
5	video_pain	0.020278
6	video_war	0.020167
7	video_death	0.018908
8	video_kill	0.018172
9	video_hate	0.017904





[18]:
```
## select top 10 keywords in the least toxic videos
lowest_toxicity_dataset = sorted_video_data.loc[lowest_toxicity_indices]

lowest_toxicity = extract_keywords_for_dataset(lowest_toxicity_dataset, num_keywords=10)

lowest_toxicity_df = pd.DataFrame(list(lowest_toxicity.items()), columns=['Metric', 'Value'])
lowest_toxicity_df
```

[18]:

	Metric	Value
0	video_musical	0.108720
1	video_music	0.107128
2	video_dance	0.105802
3	video_fun	0.105155
4	video_art	0.104730
5	video_listen	0.103442
6	video_sound	0.103164
7	video_hearing	0.102744
8	video_noise	0.102182
9	video_positive_emotion	0.025399

From our further exploration, we can answer our update research questions:

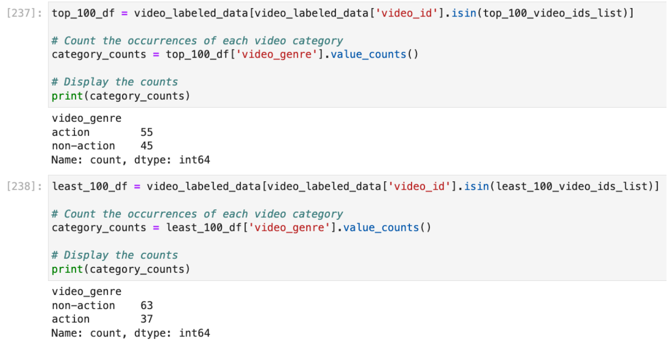
- Videos with higher toxicity are not an accurate predictor of comment toxicity. The distribution of comment toxicity appears somewhat random and may be influenced by various factors other than just the toxicity of the transcript.
- Youtubers who use more toxic language are likely to trigger more toxic comments, but it is not the only determinator.
- Surprisingly, the common topics in the most toxic and least toxic videos exhibit clear differences and patterns. Toxic videos frequently cover negative topics such as ’violence,’ ’war,’ and ’kill,’ while non-toxic videos predominantly feature positive topics like ’music,’ ’dance,’ ’art,’ and ’fun.’ Intuitively, it can be observed that the topics in toxic videos are more negative in nature compared to non-toxic videos.

From our toxicity calculation using the Perspective API, we discovered that the toxicity in the video transcript is much higher than in the comments. This difference may be attributed to the common and non-harmful use of certain language within the gaming community while playing games. Such language does not necessarily imply malicious intent, prompting us to explore methods for detecting the genuine intentions behind oral language use beyond the literal words in the gaming area.

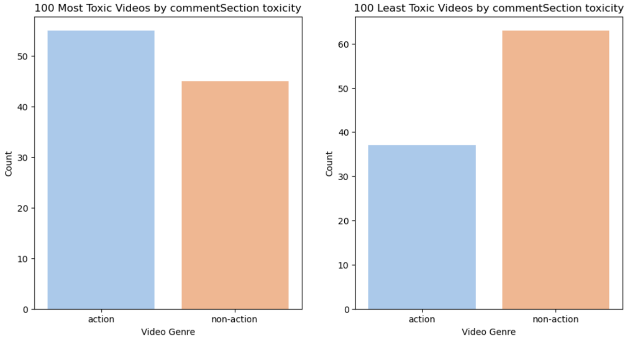
Phase 3

In phase 2, we confirmed the most common Empath keywords appear to be related with the most, and least, toxic videos. In Phase 3, we confirm its relationship with the action and non-action types.

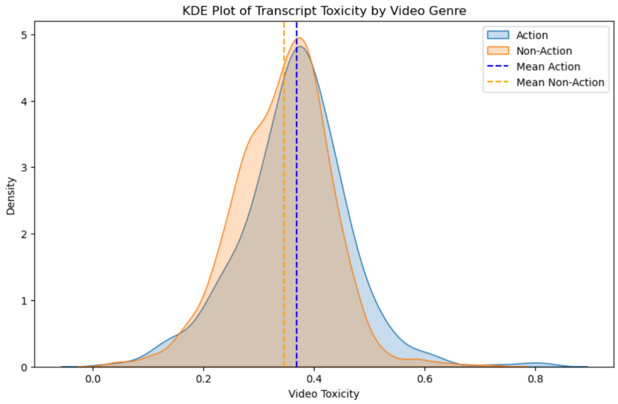
Among the top 100 most toxic videos' comment sections, 55 are action game related videos and 45 are non-action game related videos. Among the 100 least toxic videos' comment section, 63 are for non-action games and 37 are for action games. One can see that the comment sections of action games are apparently more heated than the non-action video comment sections.

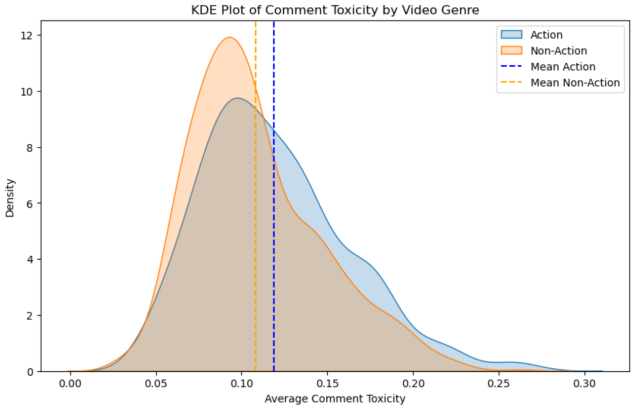


```
[237]: top_100_df = video_labeled_data[video_labeled_data['video_id'].isin(top_100_video_ids_list)]

       # Count the occurrences of each video category
       category_counts = top_100_df['video_genre'].value_counts()

       # Display the counts
       print(category_counts)

       video_genre
       action        55
       non-action    45
       Name: count, dtype: int64

[238]: least_100_df = video_labeled_data[video_labeled_data['video_id'].isin(least_100_video_ids_list)]

       # Count the occurrences of each video category
       category_counts = least_100_df['video_genre'].value_counts()

       # Display the counts
       print(category_counts)

       video_genre
       non-action    63
       action        37
       Name: count, dtype: int64
```



Following that, we created two KDE plots for both the action & non-action transcript toxicity density and the action & non-action comment toxicity comparison. We can see that Action videos contain higher average toxicity in both transcripts and comments sections.





IV. Discussion and Conclusion

In summary, we can cautiously and confidently state that our project successfully validated our two research questions:
1. Comments on action video games tend to be more toxic compared to non-action games.
2. There is a positive correlation between the toxicity of video transcripts and comments.

We were also surprised to find that the toxicity of text is much higher than that of the comments. We speculate that this may be because spoken language is more prone to toxicity than written text, influenced by YouTube's comment filtering mechanisms, and the overall positive communication atmosphere. Interestingly, despite the frequent use of profanity and offensive language in videos, the reactions in the comments section do not appear as intense as we expected. We believe this might be due to the vastly different meanings and perceived tones of words used in different contexts.

This discovery highlighted the practical value and limitations of natural language processing techniques and sentiment analysis. By understanding the impact of various factors on toxicity in the gaming community, we hope our research provides valuable insights into promoting harmonious communication in the gaming industry, contributing to its long-term development. Our future research aims to expand the study of toxicity in games to more platforms, including gaming forums and real-time communication within games.

References

Fast, E.; Chen, B.; and Bernstein, M. S. 2016. Empath: Understanding Topic Signals in Large-Scale Text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI'16. ACM.

Hutto, C.; and Gilbert, E. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1): 216–225.

Loria, S. 2018. textblob Documentation. *Release 0.15*, 2.

Perspective. n.d. Using machine learning to reduce toxicity online.

SocialBook. 2020. Top 100 gaming youtubers.

[SocialBook 2020] [Perspective n.d.] [Fast, Chen, and Bernstein 2016] [Loria 2018] [Hutto and Gilbert 2014]