Analysis of YouTube Virality with Content Categorization

Mahir Jain Anuj Sinha Sungha Kang Megan Ly

University of Washington
mahijain@uw.edu, anuj3@uw.edu, sungha24@uw.edu, lymeg000@uw.edu

Abstract

Currently, there is no way to predict the success (virality) of a YouTube video in advance. Considering the fact that more than 80 % of adults in the United States used YouTube in 2024 (Pew Research Center 2024), it is critical to understand what content does well on YouTube and come up with some form of virality score, which can then be integrated into YouTube content sorting algorithms to promote content, amongst other use cases. In addition, understanding what content does well on YouTube gives us an implicit understanding of the user base and demographics of users in an aggregated manner.

Introduction

In this project, we aim to compare the popularity of various categories of content on YouTube and analyze whether a certain category does better than others. Given that YouTube does not explicitly provide categories for a particular video, we downloaded the transcript of the selected videos and used large language models to classify the transcript into a set of preselected categories. We have also employed various normalization measures to ensure a fair comparison across videos and YouTube channels with different subscriber counts, avoiding any unfair bias that was introduced due to an unequal subscriber count.

Later, we analyze the sentiment of comments for these videos and try to see if there is any correlation between the category of videos and the overall sentiment of the comments by viewers on the video. This helped us analyze whether certain categories of videos prompt a specific reaction from their viewers.

We use common machine learning, natural language processing, and data visualization techniques in this project with the help of the *YouTubeDataApi* and *YouTubeTranscriptApi*.

Research Questions

- **RQ1**. What is the category of content that does well, and are there any noticeable trends in popularity by content category?
- **RQ2**. How can we calculate a virality score of each video, through some form of a weighted average of already existing metrics (likes, comments, views, etc)
- **RQ3**. Can we use the virality score metrics to compare the outreach of two videos?
- **RQ4**. What is the overall sentiment of comments and is there any relation to the category of content being commented on?

Data Collection

This is our data collection process and some of the assumptions and preprocessing that we did.

Selection of channels

We selected channels that are popular in their category (having more than 1M subscribers) across categories like comedy, news, and music, as well as miscellaneous categories like fitness, how-to, and talk shows. These categories are commonly found to be popular on YouTube and are of particular interest for this project (Google 2016).

The list of channels consists of 19 items - *twosetviolin, Vsauce, CNN, moneycontrol, BBC, TBS, PewDiePie, SonyMusicIndiaVEVO, wwe, smosh, Vevo, WatchMojo, Wired, BuzzFeedVideo, Vogue, howcast, fitnessblender, CinemaSins, TheEllenShow*.

While we wanted to elect channels that make cross-category content to enhance the quality of our research, finding such channels was difficult and we tried to incorporate such channels into our list as much as possible.

Transcript & Metrics Download

We downloaded data for 250 videos per channel, using the *YouTubeDataApi* and then the transcript using the *YouTubeTranscriptApi*. This was decided keeping in mind the daily request limit of 10000 for the data API and the large transcripts for each video. We also wanted to avoid any bottlenecks dealing with so much text data in the sentiment analysis as well as through the large language model for category assignment.

We were unable to download the transcript for all videos due to unavailability or restrictions and discarded all videos where the transcript could not be downloaded.

We ended up with the following number of videos for each channel -

Channel Name	Downloaded Videos
FitnessBlender	251
Smosh	251
TheEllenShow	250
CinemaSins	250
WatchMojo.com	248
WIRED	247
PewDiePie	243
British Vogue	241
BuzzFeedVideo	239
Vsauce	235
BBC	234
Howcast	226
TwoSetViolin	220
TBS	178
moneycontrol	163
CNN	151
Vevo	4

Table 1: Downloaded Videos per Channel

Sentiment Analysis on Transcript

To do a preliminary analysis of the transcript that we downloaded, we used the *Vader* sentiment analyzer to assign a polarity score to each of the videos (-1: Very Negative; +1: Very Positive) and also converted the polarity scores into buckets of positive, negative and neutral videos. We found the mean polarity score of all videos to be positive (0.7) and found more of the videos to be positive than negative, which might be related to the selection of our channels or there to be less negative content on YouTube than positive due to moderation. We also noted this skew in the data that we collected but were limited in our means to be able to find and download negative content to reduce this bias.

Category Selection or LLM Prompting

We used the following categories - *Entertainment*, *Politics*, *Lifestyle*, *Movies*, *Science*, *Sports*, *Technology*, *Finance*, *Food*, *Art*, *Nature*, *Education*, *Healthcare*, *Culture*, *Language*. We also decided to have *Other* for all the remaining categories that could not be covered in our scope. The selection of categories was based on various web sources centered around this (TutorialsPoint 2016) article. Later, we describe how we use the above list in category prediction.

In addition, to use the large language model further down in the project, we had to reduce the size of the transcript to a set threshold (max: 4000 characters). Rather than using complex text summarising techniques (Raundale and Shekhar 2021) that may cause loss of information and meaning of the content, we choose to arbitrarily select the first 4000 characters of the data or all the characters provided the original size was within the threshold. To ensure that we have not lost the meaning of the transcript and critical information, we apply the *Vader* to the shorter version of the transcript and compare it to the previous results as preliminary proof that using this arbitrary approach is valid. We find that we preserve the overall sentiment which we roughly equate to preserving the category of data. This is summarized in figure 1 below.

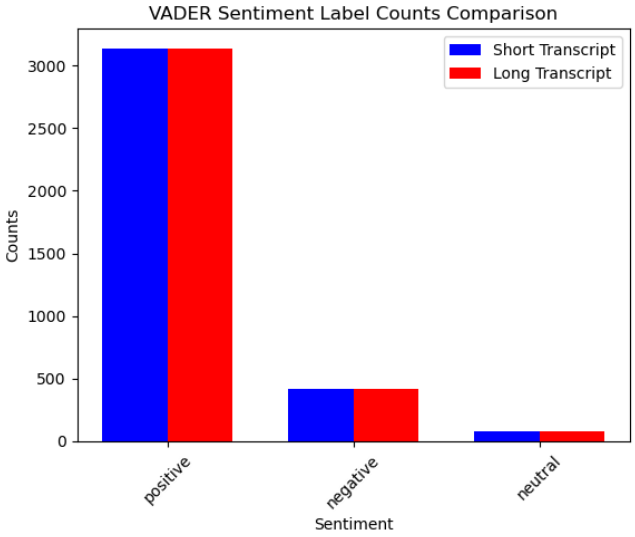

Figure 1: Vader Sentiment Labels Count Comparison

Data Analysis

We perform our primary analysis in this section, attempting to answer the research questions defined above. 2

Data Inspection

The summary statistics for the data are presented below where we look at the distributions of these statistics in figure 2. We note the skewed distribution of these metrics, in line with expectations where we will find more videos with fewer views, comments, and likes and as discussed earlier, more videos with a detected positive tone.

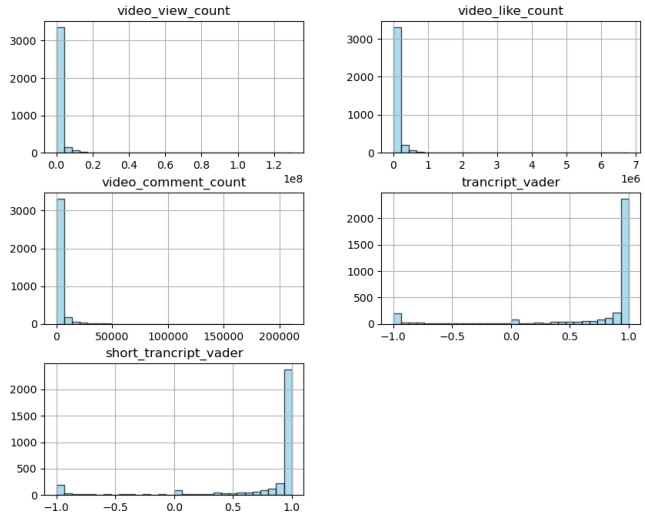

Figure 2: Frequency Distribution Of Videos Across Metrics

Transcript Categorization Using LLM

After doing, the preliminary analysis of the collected data, the next step in the pipeline was to categorize the shortened transcript using the LLM model.

Model Selection We use the LLAMA model provided by *HuggingFace* called *meta-llama/Llama-2-7b-chat-hf*. We chose this model as it is optimized for dialogue use cases. This fits perfectly in our transcript categorization use case. The selection was also based on the access and computing capacity available to us during the research.

Zero-Shot and One-Shot Prompting While using the model for categorization, the biggest challenge was to fine-tune the model using the right prompting technique. We experiment with different types of textual prompts, output formats, and zero-shot and one-shot inferencing techniques. We adopted the trial-error strategy to come up with the right sets of prompts and tested on the small subset of transcripts to sort videos in the categories list that we selected above.

We also ran the model using zero-inference, without any pre-defined list of output categories. This was for additional experimentation to see the scope of categories that LLM produces when not restricted to a pre-defined set. We used the following SYSTEM prompt.

```
You are a topic analysis expert
who is working on Social Media Data Science Research
about content-virality based on likes, dislike,
shares and number of subscriber. You always choose
the correct category for a YouTube video transcripts
from a list of categories.
```

We used the following INSTRUCTION prompt.

```
Given the following YouTube video transcript,
Classify the video into one of the categories
from the category list.

Follow these instructions:
1. Respond only single word category.
3. Respond according to the output format given below.
4. Do not add separators like '/', ',' or '\' to the
response.
5. Do not add delimiters like '\n' or '\t' to the
response.

Output Format: A single word describing the category.
```

Based on the figure 3, we observe skewness in the content categorization. This skewness arises from the limitation of our YouTube channel selection as mentioned above. We have normalized this skewness while the virality score calculation as mentioned in the later sections.

Virality analysis

Data Cleaning For the convenience of analysis, a new DataFrame was created from which only the necessary columns were extracted. Since we selected channels that own a variety of cross-category content, *Vevo*, which only has videos in the Entertainment category, was excluded from the analysis. The dataset consists of 3,628 rows and includes the following columns - channel_name, video_id, video_title, video_view_count, video_like_count, video_comment_count, filtered_category.

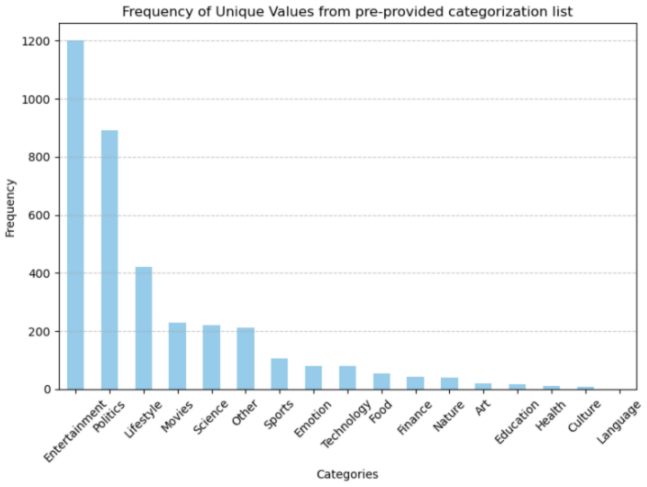

Figure 3: Category Wise Transcript Frequency Distribution

Normalization of Engagement Metrics To account for the difference in magnitudes of each metric (view, like, comment), *logarithmic transformation* was applied. This transformation enabled us to convert large numbers into small ones while maintaining the same ratio. Additionally, it reduces the data deviation, skewness, and kurtosis, and approximately conforms to normality to obtain more accurate values in analysis. (Feng et al. 2014) By normalizing the metrics and mitigating the influence of extreme values, we aimed to foster a more balanced assessment of virality. To consider the size of the channel, each metric was divided by the number of subscribers. This provided a fair comparison across videos from channels of varying sizes. Also, it allows for the assessment of the relative engagement levels of videos within each channel, considering the size of the channel's subscriber base.

$$NormalizedViewCount = \frac{\ln(ViewCount + 1)}{\ln(SubscriberCount + 1)}$$

$$NormalizedLikeCount = \frac{\ln(LikeCount + 1)}{\ln(SubscriberCount + 1)}$$

$$NormalizedCommentCount = \frac{\ln(CommentCount + 1)}{\ln(SubscriberCount + 1)}$$

Virality Score Formula Previous studies have revealed that each engagement metric contributes to virality differently. Khan and Vong (2014) examined YouTube video virality using the following factors: Video's Favorite Count, Video's View Count, Video's Comment Count, Video's Likes, and Dislikes received by a video. Research has shown that the popularity of a video is influenced not only by the metrics of the YouTube system itself, but also by network dynamics (e.g. in-links and hit counts) and offline social capital (e.g. fan base and fame), which play crucial roles in view count. They found that the variable that has the greatest impact on YouTube virality is the number of external

links (and domains) pointing to the video, and the videos that attracted more diverse domains tended to go viral faster. Among YouTube engagement metrics, the researchers found that video view counts have a controlling effect on other engagement metrics such as likes and comments. Therefore, we attempted to accurately reflect the relative importance in driving virality by assigning different levels of importance to each engagement metric. We assigned 0.5 to the number of views that we judged to most reflect external network capital and virality, and 0.25 each to the number of likes and comments.

$$\begin{aligned} ViralityScore = {} & NormalizedViewCount \times 0.5 \\ & + \quad NormalizedLikeCount \times 0.25 \\ & + \quad NormalizedCommentCount \times 0.25 \end{aligned}$$

Average Virality Score by Category Figure 4 provides a visual representation of average virality score distribution by category. Overall, the top 5 with high virality scores are identified as *Science, Art, Education, Other*, and *Emotion* categories. In particular, the virality scores of the top three, *Science, Art*, and *Education*, are almost equal. The reason why these three categories are especially popular must be found out through combination with qualitative research such as surveys and interviews. Our guess is that these three categories are popular because they can offer a wide range of content formats and inherently provoke human curiosity.

Average Virality Score by Category for Each Channel However, if you look at some of the categories with high viral scores for each channel here (Table 2), you can see a ranking that is quite different from the overall distribution.

The *Science* category still seems to be popular across various channels, but it doesn't always seem to be at the top. This is probably because the viewer demographics and preferences of each channel are different, so the categories that generate great response among each channel's audience seem to be different for each channel. Additionally, it may be influenced by factors such as video production quality and promotional efforts.

In Figure 5, we selected a subset of channels - *Howcast* and *PewDiePie* - to illustrate the distributions of average virality scores by category for each channel. The differences in viral scores among content categories within channels appear to be subtle. From the graph, it seems that the popularity of certain content does not depend solely on the channel. Rather, different characteristics inherent to each channel affect virality, resulting in varying magnitudes of viral scores for each channel. Content within a channel shows a similar viral score distribution regardless of category. Additional research seems needed that considers not only content categories and engagement metrics, but also various factors that can affect virality.

Even though it has gone through a normalization process using the number of subscribers, the normalized popularity distribution between channels shows different sizes. This suggests that unique content strategies or audience engagement metrics within the channel may be different. Therefore, an absolute comparative analysis of participation metrics seems necessary to confirm this. It seems necessary to check how data trends change by comparing absolute indicators (of total view, like, comment count) within the channel and the current normalized score analysis. In addition, acknowledging the potential bias introduced by normalization process and the limitations of normalization that oversimplifies the relationship between channel size and content performance, the development of more advanced comparison metrics seems necessary.

Sentiment analysis

Prior research did not consider video content on virality (Khan and Vong 2014). In this section, we explore the effect of a video's transcript and comment sentiment on virality in relation to the video category.

Data Cleaning We used the YouTube API to collect a sample of 50 comments from each video in the clean dataset. To analyze the sentiment of the comments, we ran *Vader* to get an average polarity score. This number was added to the data frame in an additional column. We tried a couple of methods of sampling, sorting the comments by the most recent and sorting by relevance. These methods resulted in similar results although sorting by relevance yielded slightly higher sentiment *Vader* scores. We used the latter method for the rest of our analysis since relevance more accurately captures the sentiment.

Sentiment Score by Category

From Figure 1, we know that the video transcript sentiments skew positive. Further analyzing the sentiment scores by category reveals that the language category had the highest average. When compared with the comment sentiment scores in Figure 6, we observe a significant decrease across almost all categories. Perhaps the comments have lower sentiment due to our sampling technique.

Sentiment Correlations Next, we wanted to see the correlation between sentiment and other variables.

We hypothesized that comment sentiments would be related to the video transcript (i.e. very positive videos would have very positive comments). However, there is no strong correlation when comparing the transcript *Vader* score and the comment *Vader* score.

We plotted the video transcript sentiment to the virality score. We similarly expected that content with strong emotional resonance, whether positive or negative, would elicit stronger reactions from audiences and thus higher virality. In particular, we can see that the line at +1 in Figure 7 is thicker and has a larger volume, and this clustering of more data points may mean a stronger correlation. It seems necessary to supplement through future research. We can assume for now that there is no clear correlation between virality and transcript sentiment. This suggests that factors beyond sentiment, such as content quality, relevance, and engagement strategies, may play a more significant role in determining virality.

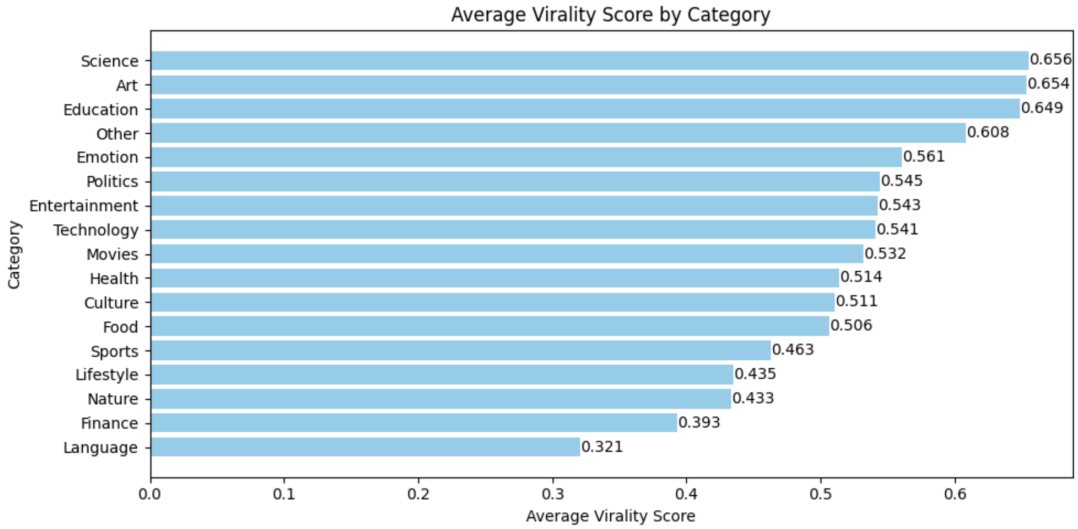

Figure 4: Average Virality Score by Category

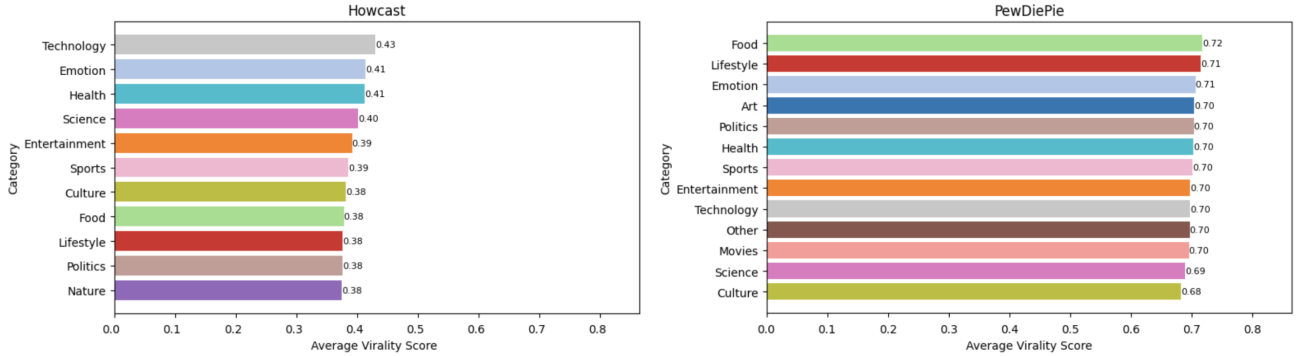

Figure 5: Average Virality Score by Category for Each Channel

Limitations and Future Work

During our research, there were a few roadblocks and limitations that we encountered. Following are a few that we would like to highlight:

- YouTubeTranscriptApi does not perform very well for very large number of videos, and we had to limit the number of videos to *5000* for this data collection step for successful execution
- Finding YouTube channels that produce cross-category content on a single channel is tough, but this would've been very interesting to analyze.
- The transcripts are not available for each video, and while we start off with a uniform number of videos to extract from each channel, we end up with a non-uniform distribution after processing.
- The overall sentiment of videos that we have collected at this stage is largely positive and while this may have something to do with our choice of channels, we do not usually expect very negative content as this stage of analysis is on the content itself and not the comments.

Our research provides insight into the dynamics of content popularity within a channel and serves as a grassroot effort to explore the various factors that can influence the virality of content within a channel. The influence of factors such as video posting time and channel characteristics, as well as content type, could be discussed in future research. In particular, it should be studied how viewer demographics, content style, and each channel's unique engagement strategy affect viral score distribution. Furthermore, studying how time affects virality (i.e. how quickly a video amasses views or likes) could be explored.

Discussion and Conclusion

Despite using a limited data sample, our study analyzing virality on YouTube through content categorization still has several implications.

In the overall virality score analysis, it was confirmed that *Science, Art, Education, Other, and Emotion* showed higher popularity than content in other categories (RQ1). The virality analysis for each channel showed a somewhat different distribution from the previous analysis, and the virality dis-

Channel Name	Top 1 Category	Top 2 Category	Top 3 Category
TwoSetViolin	Emotion	Science	Other
Vsauce	Health	Art	Food
CNN	Entertainment	Science	Politics
moneycontrol	Culture	Other	Health
BBC	Movies	Finance	Science
TBS	Emotion	Science	Movies
PewDiePie	Food	Lifestyle	Emotion
Smosh	Science	Other	Movies
WatchMojo.com	Other	Politics	Lifestyle
WIRED	Emotion	Other	Movies
BuzzFeedVideo	Technology	Emotion	Movies
British Vogue	Food	Science	Technology
Howcast	Technology	Emotion	Health
FitnessBlender	Science	Other	Emotion
CinemaSins	Other	Education	Technology
TheEllenShow	Science	Sports	Lifestyle

Table 2: Top 3 Viral Categories by Channel

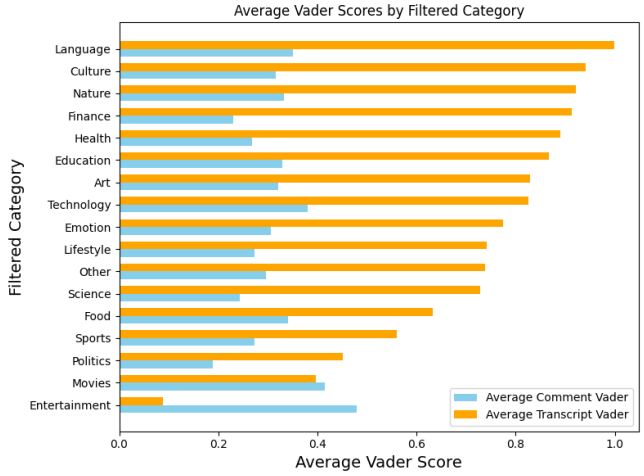

Figure 6: Comparing the comment and transcript sentiment by category.

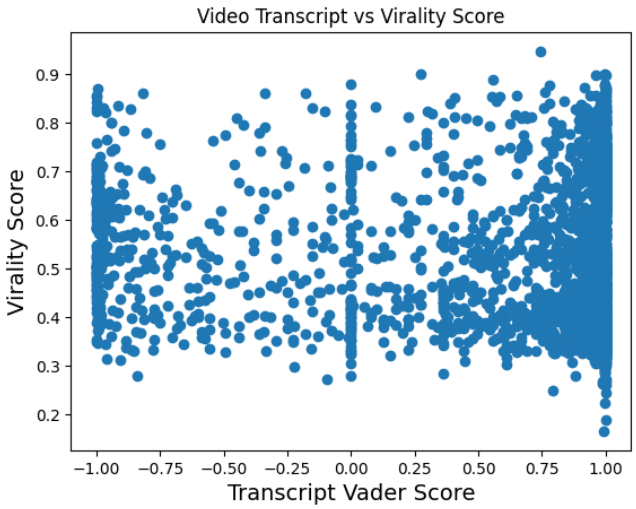

Figure 7: Transcript sentiment and virality shows no clear correlation.

tribution within each channel did not seem to have much correlation with the content category (RQ2). This could signal the fact that virality has less to do with content type but rather it is associated to the channel or creator. Since we only chose cross-cutting channels it makes sense that the *Other* category is ranked pretty high.

In addition, we expect that the analysis of viral score calculation methods will allow researchers to gain an in-depth understanding of audience participation patterns and to analyze various aspects related to user behavior within social media, such as users' online behavior, social dynamics, and the impact of digital media on society (RQ3).

Analyzing the comment sentiment did not reveal any strong relationship to the type of video category (RQ4). We expected that the comments would either be strongly positive or negative because videos that elicit strong emotions may be more viral. Further extensions of this project include sampling channels that are not cross-content to see if sentiment still remains positive and using .

Overall, our research is useful for many stakeholders. By analyzing the sentiment of comments and virality scores of videos, YouTube, as a platform, can improve its recommendation algorithms. For creators, understanding what kinds of content their users like best can also boost their channel's growth. We hope that in the future, there can be more research done in this area to better understand the nature of social media platforms like YouTube.

References

Feng, C.; Wang, H.; Lu, N.; Chen, T.; He, H.; Lu, Y.; and Tu, X. 2014. Log-transformation and its implications for data analysis. *Shanghai Archives of Psychiatry*, 26(2): 105–109.

Google. 2016. Consume Insights. https://www.thinkwithgoogle.com/consumer-insights/consumer-trends/top-content-categories-youtube/. Accessed: 2024-03-01.

Khan, G. F.; and Vong, S. 2014. Virality over YouTube: an empirical analysis. *Internet Research*, 24(5): 629–647.

Pew Research Center. 2024. Social Media Fact Sheet. https://www.pewresearch.org/internet/fact-sheet/social-media/#:~:text=the%20same%20platform.-,Which%20social%20media%20platforms%20are%20most%20common%3F,most%2Dwidely%20used%20online%20platforms. Accessed: 2024-03-01.

Raundale, P.; and Shekhar, H. 2021. Analytical study of Text Summarization Techniques. In *2021 Asian Conference on Innovation in Technology (ASIANCON)*, 1–4.

TutorialsPoint. 2016. What are the different types of YouTube Video Categories? Accessed: 2024-03-01.