New York Institute
of Technology

New York Institute of Technology
Department of Computer Science
CSCI 436 M01/W01
DTSC 701-W01D
Big Data Analytics Project Report
Fall 2020 Semester
Prof. Xueqing Huang

Engaging YouTube Communities
Analysis of Global Trending Youtube Videos

Christopher Griffith
Helen
Jennifer
Greg

ABSTRACT

The focus of this project was to improve the YouTube algorithm for determining trending videos. In this work, comment counts are used to determine community engagement per video for each region. The means of comment counts per region are utilized as thresholds to classify each video as either engaged or non-engaged. Videos with high community engagement should be considered trending videos.

I. INTRODUCTION

YouTube, an online platform that allows users to share videos, was created in 2005, and bought over by Google in 2006. It currently has over a billion users daily, millions of content creators on the platform, and continues to thrive and be successful. Not only have the creators of YouTube achieve financial success due to the platform's success, but also many content creators on the platform have made considerable financial gains from the success of their content on the platform. However, there are many factors that influence the success of a creator's content aside from just the content itself, thus leading our curiosities to understanding what factors influence and makes a YouTube video trend.

We are interested in improving the YouTube algorithm for determining trending videos. We have noticed that some of YouTube's trending videos do not have a lot of views, likes, dislikes, or comments. We believe that in order for a video to be considered trending, it should have a highly engaging community, specifically in terms of comment counts. The idea behind this is that, when viewers watch a video, viewers may not necessarily engage with that video. They may just view it, and that's it. However, when we view a video that piques our interest, we are inclined to like or to dislike that video, and to leave a comment on that video. When multiple viewers are inclined to express their thoughts and viewpoints of that video through commenting, we believe that these videos should be considered trending. If a video has a highly engaged community, this video should be recommended as a trending video.

We are proposing to improve YouTube's trending videos algorithm because YouTube is a profitable platform for content creators, advertisers, sponsors, and YouTube itself. By improving the trending videos algorithm, we believe that ads, contents, and sponsorships can be better targeted to the corresponding demographic of the community based on community engagement, and more specifically, based on contents of the comments in their respective videos.

II. RELATED WORKS

Comment counts in trending videos. YouTube trending videos represent content that targets viewers' attention over a relatively short period of time and has potential of becoming popular [1]. After uploading, trending videos are declared as such within just several hours. For trending videos, about one half of videos receive less than a few hundred comments per video whereas 30% of trending videos receive more than 1,000 comments. This behavior may be attributed to the level of popularity of trending videos.

Comment counts with respect to days to trending, likes, dislikes, and views in trending videos. Comment counts are generally linearly related to views, likes, and dislikes [2]. YouTube generates huge knowledge, knowledge that is unstructured, semi-structured, and unpredictable in nature, through its community feedback, which includes comments on videos, range of likes or dislikes, and the range of subscribers for a specific channel. Analyzing these knowledge points allows mining for getting implicit data, such as uses, videos, classes, and community classes. Most businesses and content creators will look for these kinds of feedback after launching their products, which are crucial to understanding customers' sentiments regarding their products or services.

III. SYSTEM MODEL

The proposed trending videos recommendation system is an enhanced model that classifies trending videos as such if they have a highly engaged community, specifically in terms of comment counts. We pulled data from Kaggle as well as YouTube Data API. The data is processed and a threshold is determined for what is considered an engaging video. We then train the binary classification model, find the recommendations, and then evaluate how well our model performed based on the given data.

Phase I: Data Collection: Data was collected from a Kaggle dataset, which originated from YouTube Data API. The training data consisted of over 375,000 records, with 200 trending videos per day for 6 months, ranging from the fourth quarter of 2018 to the first quarter of 2019 for 10 regions. The test data was from the YouTube Data API, with 200 videos per day for a span of 3 days, ranging from December 10, 2020 to December 12, 2020.

Phase II: Data Processing: After collecting the data, we needed to process the data. We removed any videos that had views of less than 100,000, videos with no comments, videos with no likes, or videos with no dislikes. By removing any videos that had no comments, or that did not allow for comments, we removed any videos that did not allow for community engagement. For certain regions, which were the United States, India, United Kingdom, and Canada, our removal criterion removed about 20% of videos. On the other hand, for countries such as Japan, Russia, and Mexico, our removal criterion was not too ideal as we removed over 50% of the videos [Figure 1]. This could be due to video restrictions, region restrictions, or region customs in terms of how engaged online communities are, how well the community communicates with each other, and how expressive they are. The same removal criterion was applied to both the training and test datasets.

Feature Analysis: The correlation heatmap of the training data compares the views, likes, dislikes, and comment counts per region [Figure 2]. We can see the habits of each region in terms of comment counts to their respective views, likes, and dislikes. For instance, there is a very high correlation of 0.94 between comment counts and likes for videos in Japan. This shows that viewers who clicked the like button were very likely to leave a comment as well. On the other hand, in the United Kingdom, there was a low correlation between views and comments. This shows that although viewers watched a video, they were not likely to leave a comment as well.

The correlation heatmap of the test data also compared comment counts to views, likes, and dislikes by region [Figure 3]. Similar to the training data results, we see that there is a high correlation of 0.90 between comment counts and likes for videos in Japan.

Region Region Code	Records	Records Removed	% of Removed Records
United States (US)	40949	6220	15.189626%
India (IN)	37352	8955	23.974619%
United Kingdom (GB)	38916	5886	15.124884%
Japan (JP)	20523	13990	68.167421%
France (FR)	40724	23987	58.901384%
Mexico (MX)	40451	26111	64.549702%
South Korea (KR)	34567	18025	52.145109%
Russia (RU)	40739	25885	63.538623%
Germany (DE)	40840	20329	49.777179%
Canada (CA)	40881	8359	20.447151%

Figure 1: Percentage of removed records per region after applying removal criterion

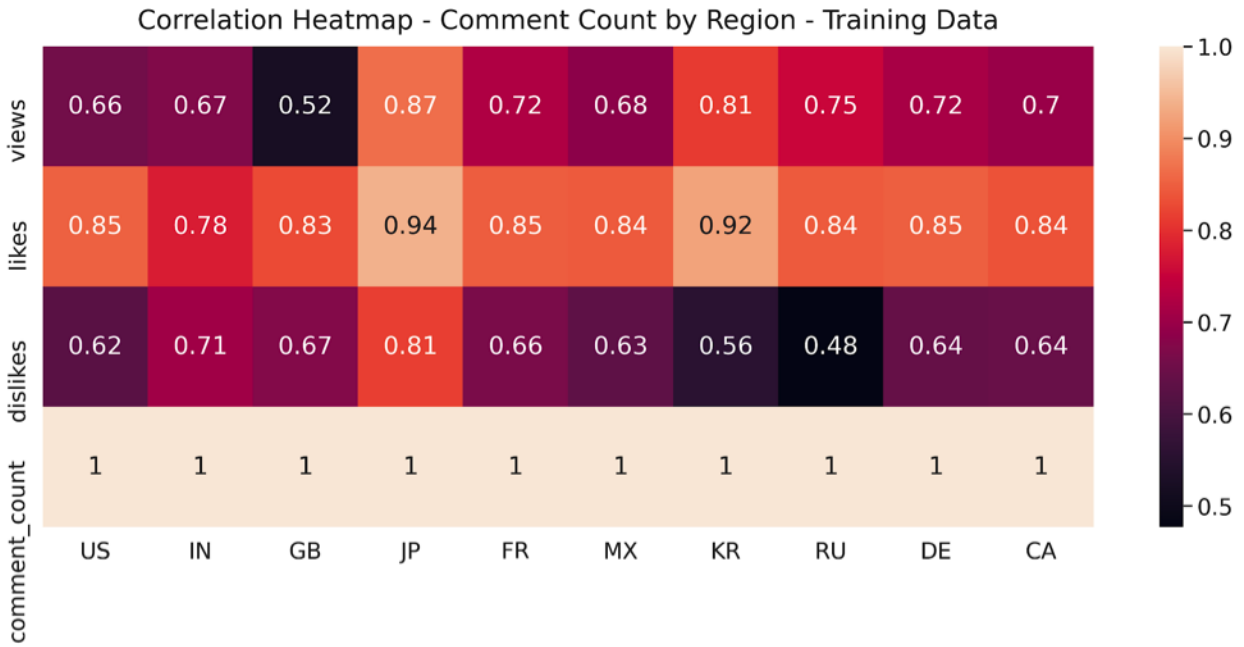

Figure 2: Correlation heatmap of comment count by region for training data

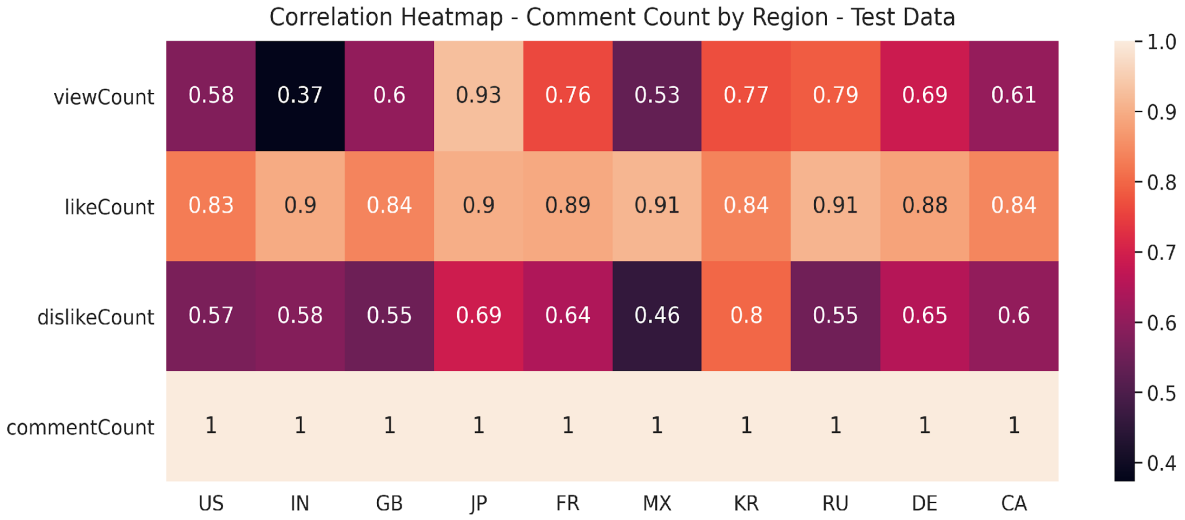

Figure 3: Correlation heatmap of comment count by region for test data

Phase III: Recommendation System: For our model, we used a binary classification machine learning model. Specifically, we used a decision tree algorithm in order for us to determine which videos were considered engaging, and which videos were non-engaging with respect to their comment counts. The parameters that we used from each dataset for each record were views, likes, and dislikes. The decision tree algorithm was then able to determine if those videos were considered engaging or non-engaging.

We used a threshold count for all of the data within each region based off of the mean of the comment count. The boxplots show the comment counts for every video within each region [Figures 4, 5]. We can see that there is a huge amount of outliers within each region for both the training and test datasets, which means that there are videos that have substantially high amounts of comment counts compared to the rest of the videos in that subset. As a result, we hashed the data, removing all of the outliers [Figures 6, 7]. Removing outliers was beneficial as the spread of comment counts reduced from up to 1.2 million comments down to approximately 25,000 comments maximum. We originally further removed outliers again, but did not see a significant difference in the spread. Additionally, if we continued to remove outliers, we may end up working with a comment count mean that doesn't account for much. As a result, we settled with the means that were achieved from removing the outliers once for every region. These means were utilized as thresholds. For videos that had comment counts above the mean, they were considered to be engaging videos, and for videos that had comment counts below the mean, they were considered to be non-engaging videos.

We then trained our model using three different sets of training and test data. We first trained the model using an 80/20 split of the training data, where we trained the model with 80% of the training data, and then tested the model with the other 20%. We then re-trained the model using 80% of the training data, and testing it with our test dataset from December 10, 2020 to December 12, 2020. Finally, we trained the data again with 100% of the training data, and tested the data with our test dataset.

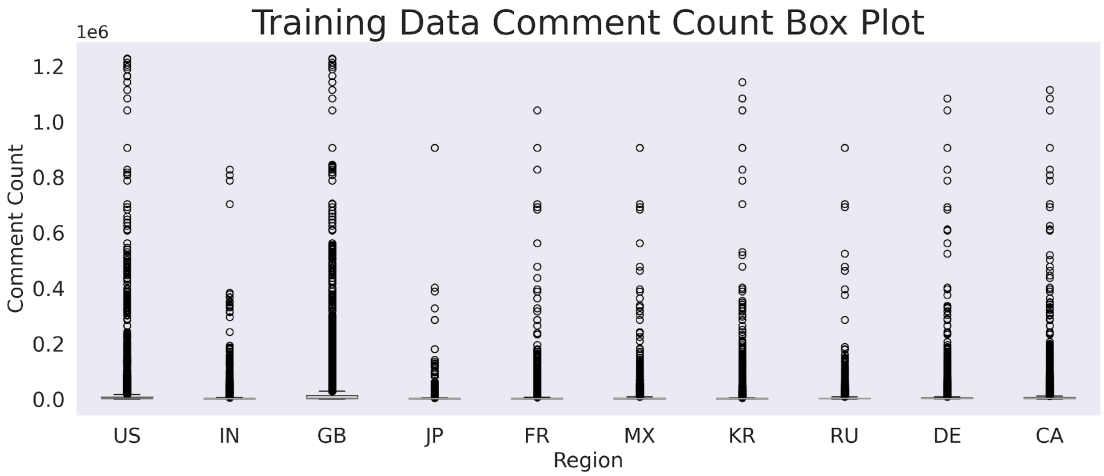

Figure 4: Comment count box plot for training data

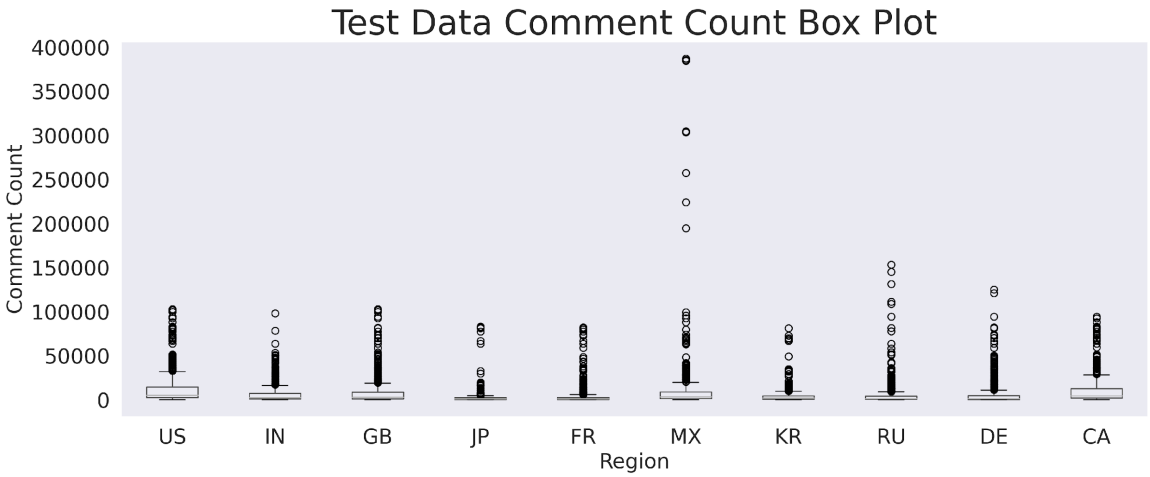

Figure 5: Comment count box plot for test data

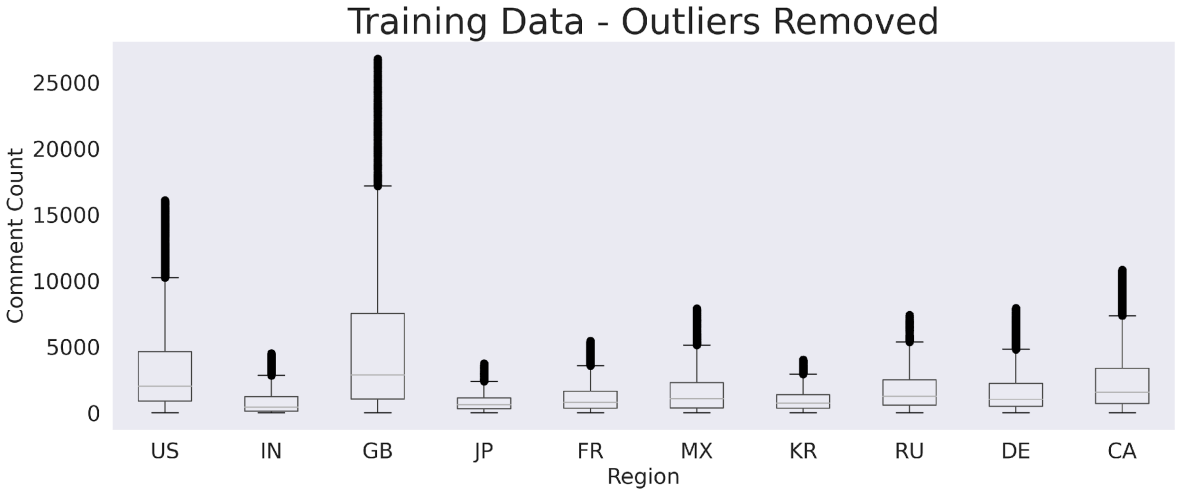

Figure 6: Comment count box plot with outliers removed for training data

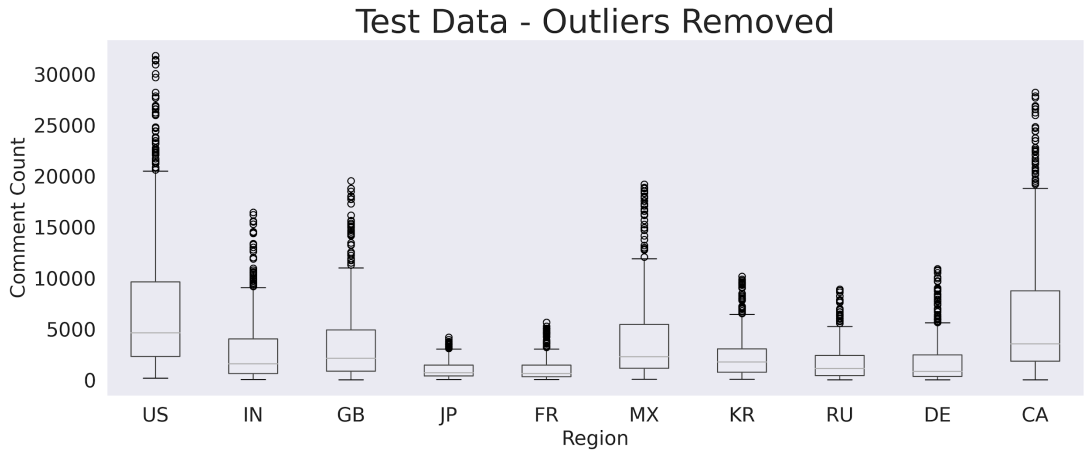

Figure 7: Comment count box plot with outliers removed for test data

IV. ANALYSIS AND DISCUSSIONS

For the first training, we used an 80/20 split of the training data for both training and testing. We can see the corresponding accuracies of those decision trees [Figure 8], which depicts that model was relatively accurate. The most accurate is for the United Kingdom region at 91.1202% accuracy. For every other region, we achieved an accuracy of at least 82%.

For the second training, we used 80% of the training dataset for training, and then used our test dataset for testing. The results for this were less ideal as we see that a lot of regions were incorrectly classified. Regions that had low accuracy are depicted with red [Figure 9], such as India with 56.67% accuracy and South Korea with 51.08% accuracy. For the regions depicted with blue, the model performed relatively well. We notice that, for the red regions, they are similar to the regions from when we removed a lot of the records [Figure 1] during data processing. Therefore, we suggest that it is possible to improve the performance of our model if we increase the number of records in our training model, with the exception of Canada, which seemed to have done poorly with respect to our model despite having a decent amount of training records.

Finally, we trained the model one last time with 100% of the training data and tested the model using our test dataset. We achieved similar results as our previous model, except all of the accuracies decreased [Figure 10]. This could be a result of greediness as we used all of the training data, causing overfitting to occur.

When we trained and tested our model using the training dataset only, we produced ideal results with high accuracy. However, when we used the test dataset on our model, we did not achieve ideal results for every region. For instance, India initially had a high accuracy of 89.1882% when we used only the training dataset. However, accuracy decreased to 57.0611% when we used the test data. Similarly, in the correlation heatmap for the training data [Figure 2], there was a correlation of 0.67 between comment counts and views for India. However, in the test data, the correlation between the comment counts and views was only 0.37 [Figure 3]. The significant difference in accuracy and correlations may be due to the vast time difference of the datasets. We

trained our model using data from 2018 and 2019, whereas our test data is from December 2020. The discrepancy could be because we are testing current data with a model that was trained with and worked for relatively older data. It is also possible that due to cultural reasons, there was a shift in online video engagement practices between 2018 to 2019 and 2020. Additionally, recall that during data processing, we removed 23.97% of records. The discrepancy in accuracy may be because we did not have enough dataset as our test data was only over the span of 3 days.

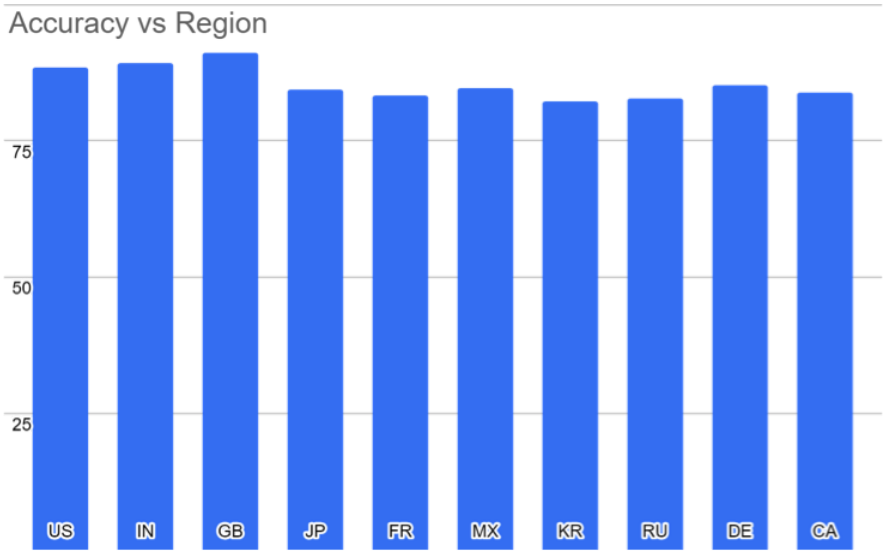

Figure 8: Accuracy vs. region with 80/20 split of the training data for both training and testing

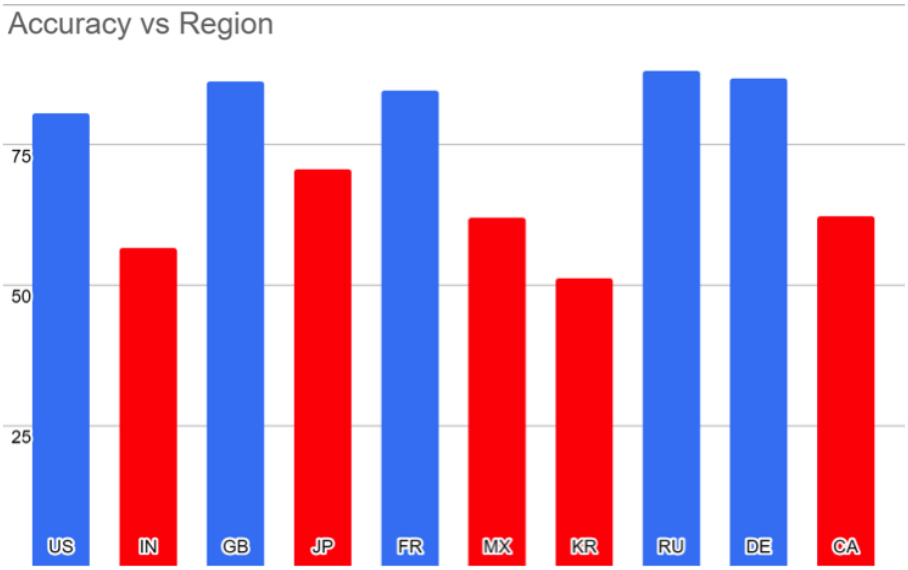

Figure 9: Accuracy vs. region with 80% of the training data and testing with test data

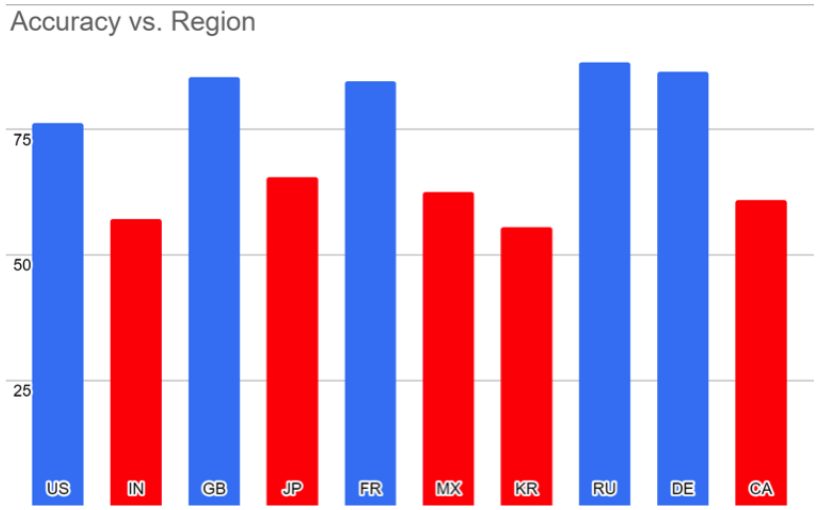

Figure 10: Accuracy vs. region with 100% of the training data and testing with test data

V. CONCLUSIONS

In our model, comment counts were used as a threshold, and parameters of likes, dislikes, and views were used to determine whether videos were considered to be engaging or non-engaging. We used the comment count means per region as the threshold. Once we determined the means for each region, we trained the data, and then classified each video as either highly engaged or lowly engaged, based on the mean. Videos above the comment count mean were classified with a '1', and videos below the mean were classified with a '0'. For our model, more data may not necessarily be the best. Rather, it requires more processing and further tuning to identify possible cultural differences between each region.

YouTube videos with high community engagement and relatively lower views can be considered as trending videos. Viewership does not guarantee an engaging community. It is important to note that how engaging a community is depends on the region. We saw that some trending videos have comment counts that were below the mean. Approximately 70% of the trending videos had comment counts at or below the mean, and the rest of the videos had comments counts above the mean, which we determined to be highly engaged videos.

Profitability for content creators, advertisers, and sponsors is not solely based on viewership. An established engaging community is also an important factor. If viewers are highly engaging with a video, specifically through comment counts, the video should be considered trending. This allows the video to reach a greater number of audiences. Consequently, based on comment contents, the content creators, advertisers, and sponsors, can better target their content to the corresponding demographic of the community.

VI. FUTURE WORK

For trending videos with high comment counts, we would like to analyze the text information to see what those comments were about. The contents of the comments would provide valuable

information to the content creators and advertisers, as they would provide insight as to what viewers are interested about in regards to that video's contents, and what is it that inclined the viewers to comment. This could inspire the content of the next video of that content creator or advertiser. We would also like to analyze the text information of the comments, as well as the description of the videos, and their impact on the video's viewership.

We would also like to improve our model to account for more up-to-date data, and consequently, to improve accuracy. As our model was trained and tested in an offline setting, we would like to see how our model would perform in an online setting. To improve performance, we would also like to improve our model to be able to account for regional differences in terms of online community practices. Additionally, we would explore other methods of visualizing our data.

VII. CONTRIBUTIONS

Christopher Griffith

- Implement the YouTube Data API to gather data on the latest trending videos
- Cleaned the training and test data using the proposed removal criterion
- Applied the binary classification model to the test data for YouTube Data API
- Analyzed the results of the model and provided rationale for any discrepancies
- Contributed to the development of both the presentation and project paper

Helen -

- Researched related works on YouTube trending videos and comment counts
- Worked on the presentation
- Wrote and completed the project paper

Jennifer -

- Presentation formation
- Research

Greg -

- Research
- Implemented Data Visualization
- Assisted with Presentation

REFERENCES

[1] I. Barjasteh, Y. Liu, and H. Radha, "Trending Videos: Measurement and Analysis," arXiv: 1409.7733, 2014. [Online]' Available: https://arxiv.org/ftp/arxiv/papers/1409/1409.7733.pdf

[2] S. Amudha, V R. Niveditha, Dr. P.S. Raja Kumar, M. Revathi, Dr. S. Radha Rammohan, "Youtube Trending Video Metadata Analysis Using Machine Learning," in *International Journal of Advanced Science and Technology*, vol. 29, no. 7s, pp. 3028-3037, May 2020.