**Trending YouTube Video Analysis & Prediction**

Shun-Ting Wang  
 University of Maryland  
 stw@umd.edu

Kristian Nicholas Koeser  
 University of Maryland  
 krisnk@umd.edu

**1 INTRODUCTION**

YouTube is the most well-known online video streaming platform that was founded on February 14, 2005. YouTube has more than 2.5 billion monthly users who collectively watch more than one billion hours of videos each day. It also has an unprecedented social impact, influencing popular culture, internet trends, and creating multimillionaire celebrities. Our intention for this project is to analyze what the trending videos had in common, what the differences of the trending videos between different countries are, and what elements and directions a video creator could do to make the video more likely to be trending.

The dataset includes several months of data on daily trending YouTube videos. The date ranges from November 2017 to June 2018. Data is included for the US, GB, DE, CA, FR, KR, JP, MX, RU, and IN regions (United States, Great Britain, Germany, Canada, France, Korea, Japan, Mexico, Russia, India, respectively), with up to 200 listed trending videos per day. Each region’s data is in a separate csv file. Data includes the video title, channel title, publish time, tags, views, likes and dislikes, description, and comment count. The data also includes a category\_id field, which varies between regions. The whole dataset has 750145 data. US has 40950, Russia has 40739, Mexico has 40451, Korea has 34567, Japan has 20523, India has 37352, Great Britain has 38916, France has 40724, Germany has 40840, and Canada has 40881 data. The dataset has 16 variables including video ID, trending date, video title, channel title, category ID, publish time, tags, views, likes, dislikes, comment count, thumbnail link, comments disabled, ratings disabled, video error or removed, and description.

While performing our analyses for data of the United States videos, we decided to create two regression models. The first was made to predict how many days consecutively a video would trend based on its metrics on the first day it started trending. The second allows for the prediction of the number of dislikes on a trending video based on its likes, comment count, views, and days trending since YouTube no longer allows the public to see the number of dislikes on videos. We also analyzed the whole dataset which includes the data of ten countries. In this analysis, we want to answer the following questions: How long can a video trend in each country? Are there videos that trend in more than one country at once? What’s the top video category for each country?

**2 LITERATURE REVIEW**

YouTube as a platform has been researched previously because it is the most iconic user generated content platform. However, there has been less analytical work done on the trending videos specifically. We want to examine more of the cultural differences between countries by looking into the top categories and see how strong the correlations of the metrics of the videos are.

Gayakwad et al. concluded that knowing the best timing to post a video on YouTube is not enough to make the video trending. Some other factors such as titles, thumbnails, tags,and the number of the subscribers are factors that a creator also needs to consider. [1]

Zhou et al. studied the impact of YouTube recommendation system on video views and found a strong correlation between the view count of a video and the average view count of its top referred videos, meaning that YouTube recommendation helps viewers discover more videos of their interest rather than the trending videos only. In this case, we can see the cultural differences between each country and how YouTube impacts the popular culture.[2]

In another research, Brodersen et al. used a large database of video and new measures to quantify their popularity distribution across different geographic regions and found out that YouTube videos exhibit strong geographic locality of interest.[3]

For the predictive model, Niture implemented a simple Logarithmic Linear Regression algorithm to predict the number of views for a YouTube trending video and concluded that the basic linear regression model gives a moderate accuracy percentage of 77.3%. [4]

Lups, a-Tătaru et al. confirmed that in order to engage people when they watch a YouTube channel for the first time, to create a trailer, which is similar to the function of movie trailers, would be the best strategy. By doing so, it can minimize the chance that people will leave to watch another channel’s videos. This finding coincides with our finding from wordcloud. “Trailer” is the top title keywords of the trending video. [5]

**3 METHODOLOGY**

Analysis for this project was conducted on the United States video data alone as well as the data from each available country.. Predictive models were then built on our analysis. We will use Python with packages like Pandas, Matplotlib, and Seaborn to analyze and visualize a dataset that was collected over 205 days and use Python’s statsmodels module to implement Ordinary Least Squares(OLS) method of linear regression to create our predictive models.

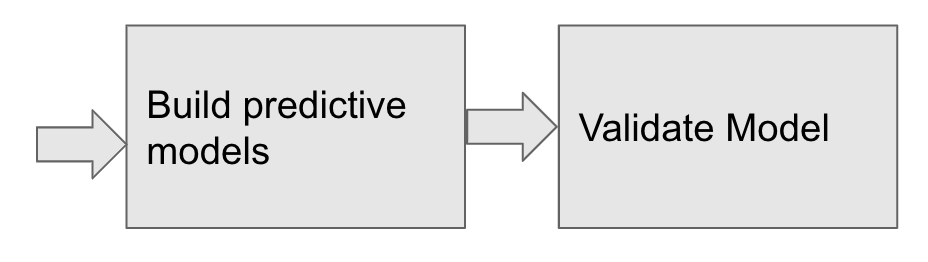
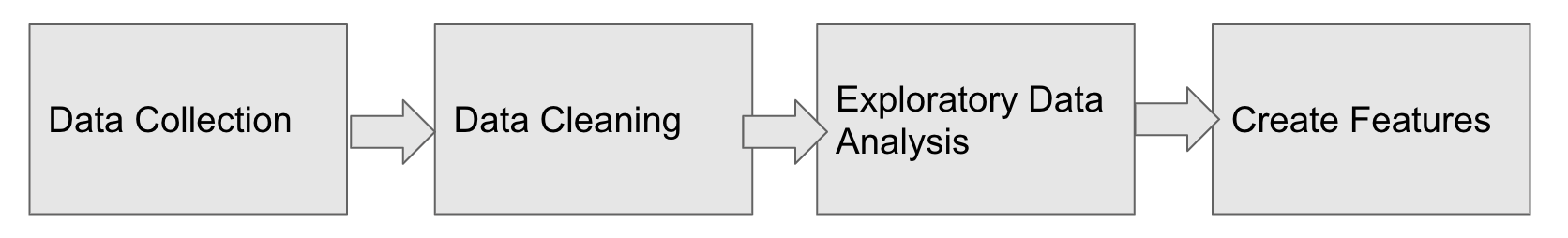
****

Figure 1: **The pipeline of the project**

3.1 **Analyzing the basic statistics of trending videos in United States**

The data was obtained from Kaggle. Since we have around 750,000 rows of data, we started our analysis with the United States video dataset. To clean the data we had to convert the date published and date trending values to the pandas date time format as well as labeled each of the videos with a days trending column which would display how many times the video recurred in the dataset which essentially told us how many days the video was on the trending page. To better suit our end goal, we created a column that stores the day of trending of each video. With this feature, we built our predictive model based on this value along with like, dislike, and comment counts.

3.2 **Analyzing the basic statistics of trending videos in 10 countries**

We dropped the null values which we found a lot in the Japan, Korea, Mexico, Russia, and India data. After cleaning the data, we made some graphs and charts to compare different metrics and to see the correlation of the video features. We also had to map the country code to the whole dataset since we concatenated all the data from each country.

3.3 **Building the predictive model based on the analysis of trending videos in United States**

After having the new feature, days trending, created we built the model to predict the number of dislikes since YouTube had officially hidden the dislike number from the public. Only the video owner can see the statistic. Therefore, making a predictive model would be really insightful for people who wonder about the dislikes number of a YouTube video.

**4 EXPERIMENT RESULT**

4.1**Correlation between basic video metrics**

For many of the analyses it was decided to work only with the United States data set due to the overall size of the combined data as well as difficulties with cleaning some aspects of the data from the other countries. The United States data was also chosen as a focus due to its familiarity. Initial analyses of the data focused on determining how each of the video’s four visible metrics relate to each other. The metrics that are present in the data set include: likes, dislikes, comment count, and views. Another avenue of examination was the video title length and determining if it has any impact on views. In order to quantify these relationships scatter plots were created and the Pearson correlation coefficient was calculated for these relationships as well. It was determined that all of the video’s metrics had positive correlation with regards to both views and comment count and that the metrics with the highest correlation are likes and views as well as likes and comment count.

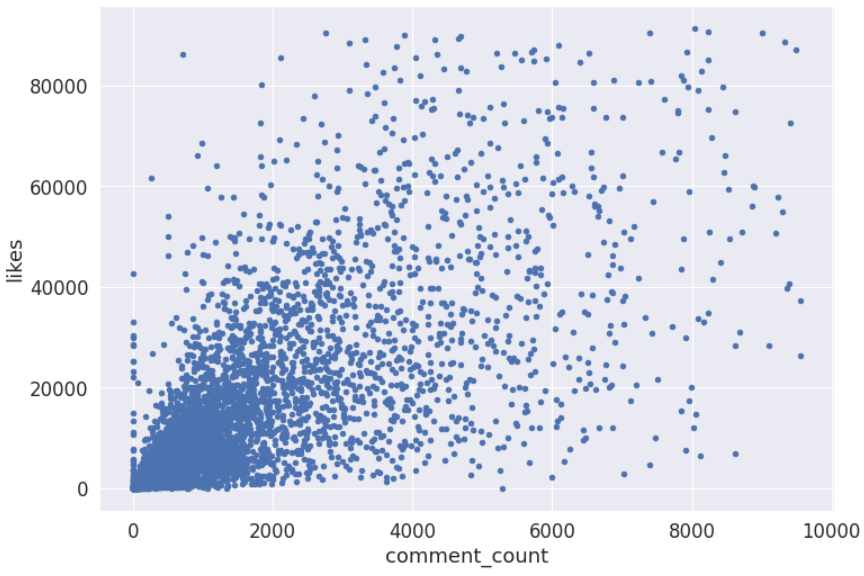
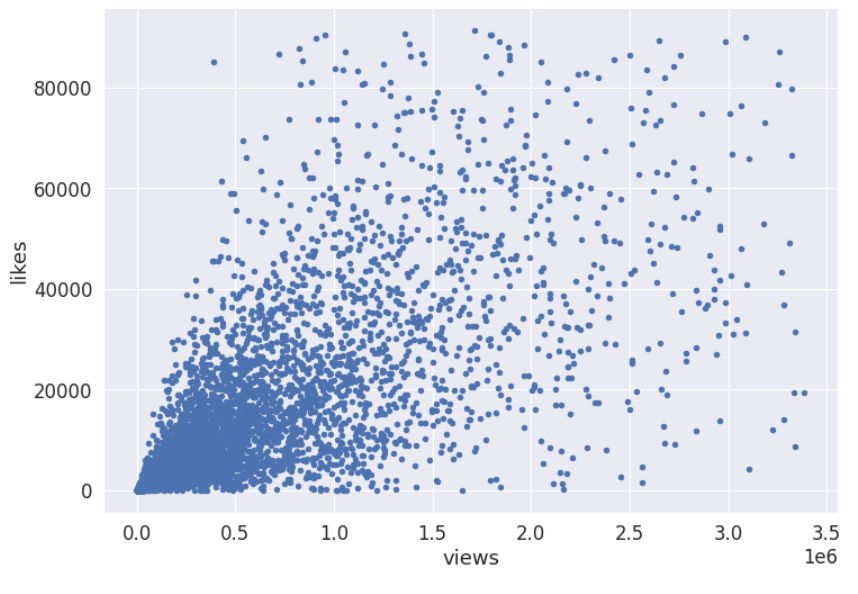


Figure 2. Scatter plots showing the relationship between views and likes (left) and comment count and likes (right).

4.2**Wordcloud**

It was also found that video title length varies greatly in this data set and has no correlation to views. In order to examine common words in the titles of the trending videos, word clouds were generated. One word cloud was generated using the entire United States data set and additionally a word cloud was generated for each of the individual categories. Several words and phrases appeared dominantly in the data set with “New”, “Official Trailer”, and “Official Video” being most common across all of the trending videos. This is in part explained by the fact that Entertainment is the category with the largest number of videos in the United States dataset.



Figure 3. Word cloud showing the frequency of words in the titles of the trending videos in the United States dataset with the size of the word increasing with the number of instances of the word in the dataset.

4.3**Publish day analysis**

Additionally, the day of the week that each video was published was determined and the videos were counted and visualized using this metric in order to see if there were any apparent trends between videos that are present on the trending page and the day they were published. With regards to the publish day of the week, it was found that each weekday, Monday through Friday, are about even in terms of representation on the trending page but that it was about half as common for a video published on Saturday or Sunday to go trending.

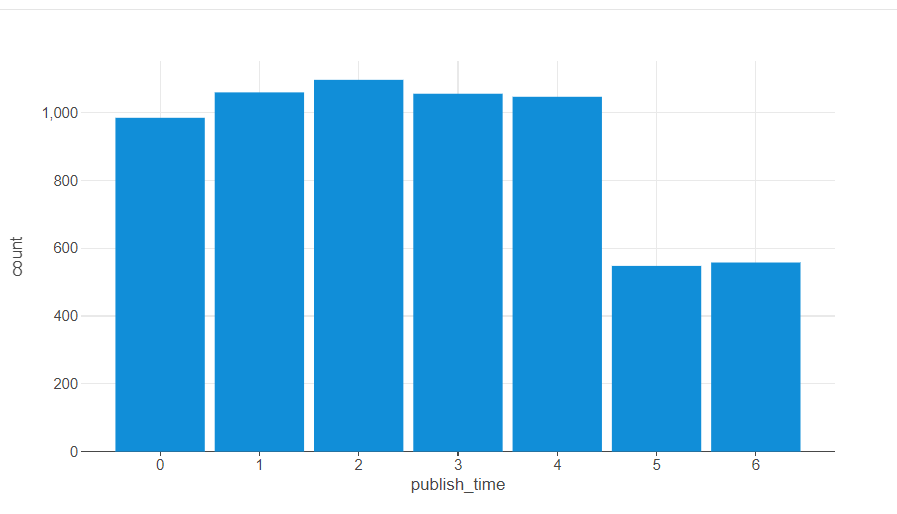


Figure 3. Chart showing the count of videos published on each day of the week with Monday being 0 and Sunday being 6.

4.4**Worldwide continuous days trending and category analysis**

Each individual country’s data set was then combined to examine the most popular categories of videos across the world as well as which countries video’s trended for the longest periods of time. Great Britain had the longest trending videos with a video that trended for 37 days while Mexico had the fewest days trending for videos with no video trending for more than 4 days. As a result of this, countries with videos that trended for longer had fewer unique trending videos within the timeframe of the dataset. However it is noted that several videos trend in multiple countries within the same time frame. Across all of the available data, the most frequent categories for trending videos were: Entertainment, People & Blog, News & Politics, Sports, Comedy, and Music.

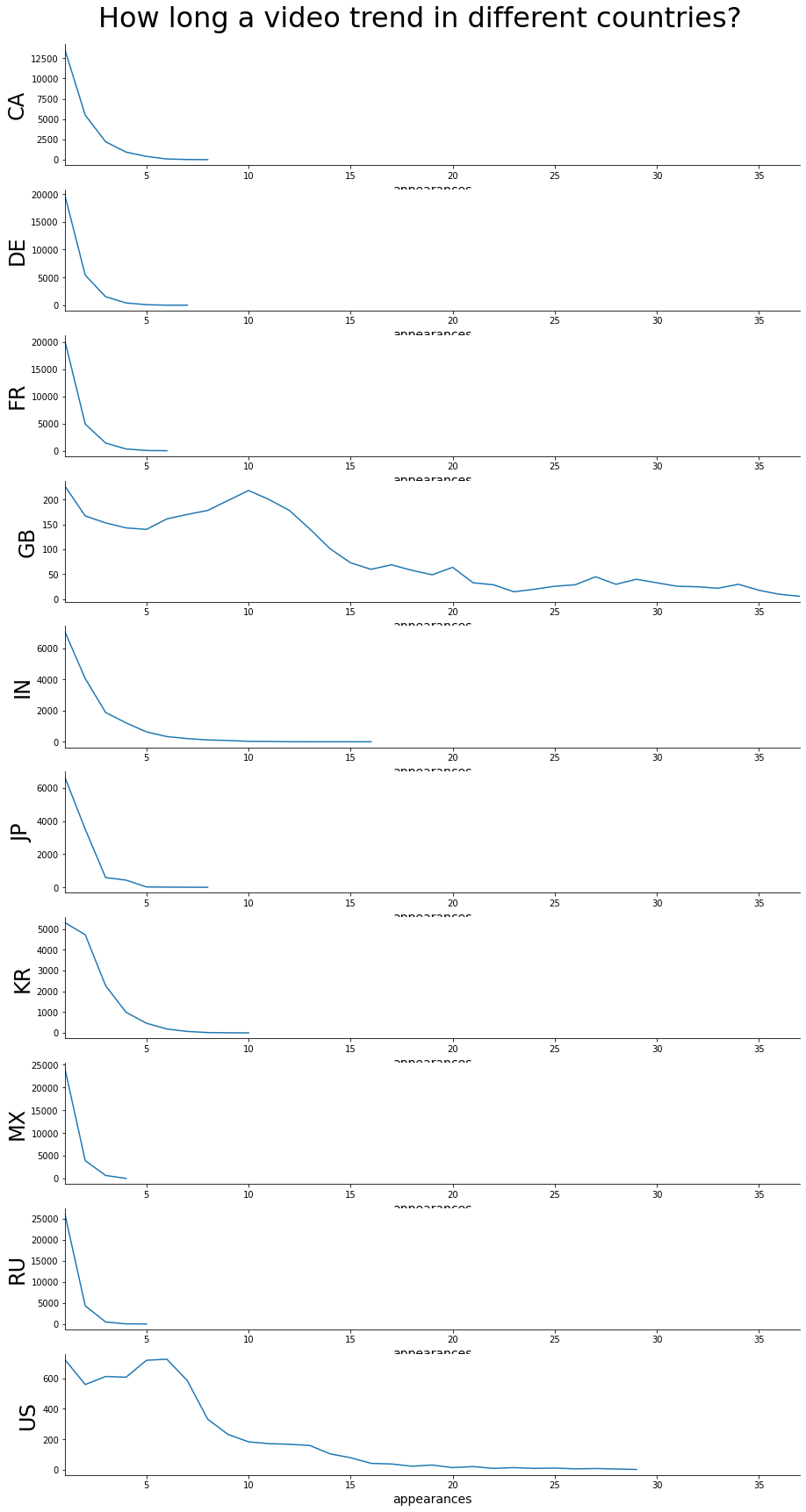
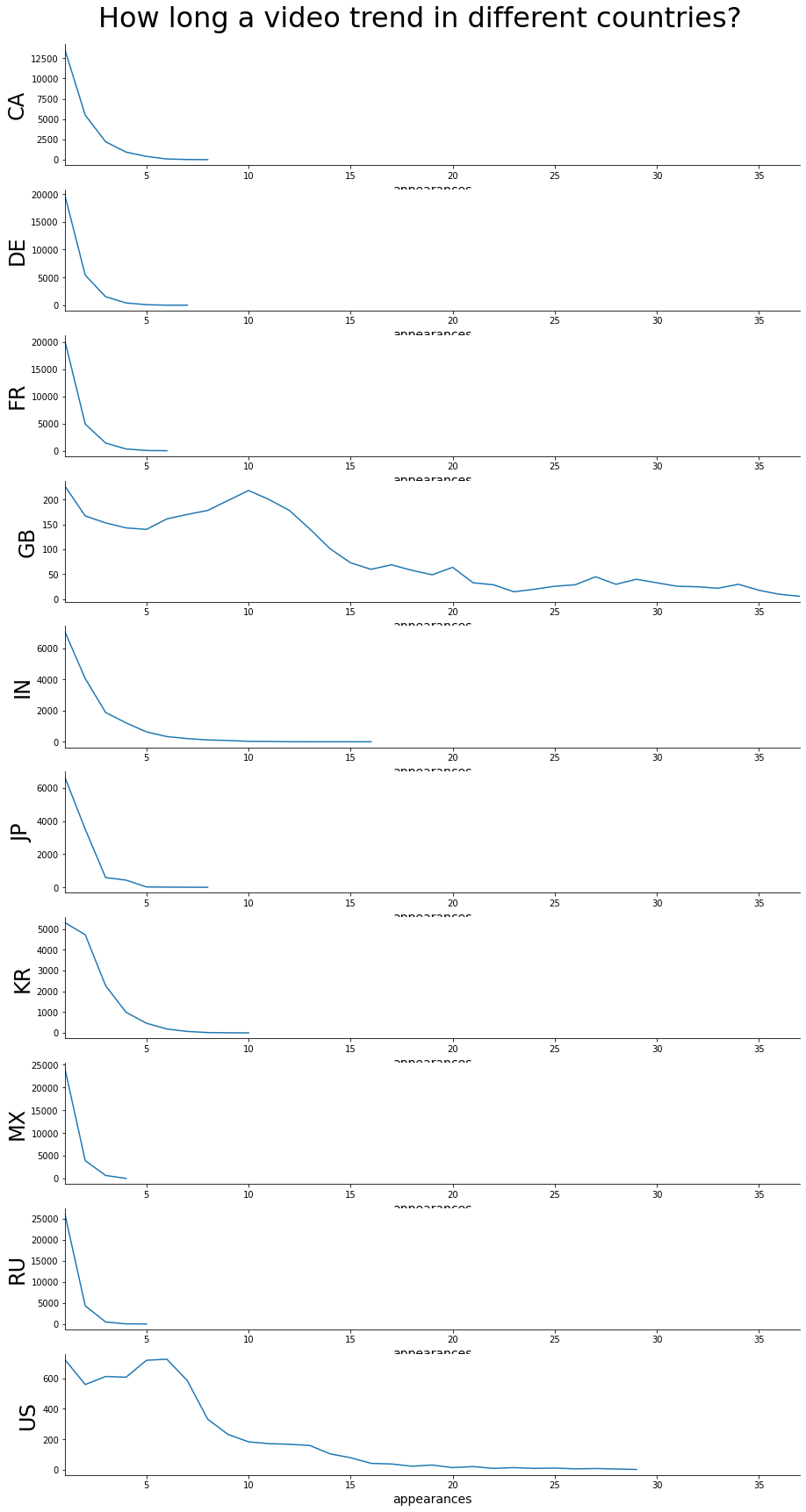


Figure 4. Line charts showing the count of trending videos by days trending across all of the countries in the data set.

4.5**Predictive model**

As part of the analysis of the data set, two multiple linear regression models were created. Both of these models utilized the default values of the Python statsmodels.api. This includes a confidence interval of 0.975. The first model predicts the total number of days that a video would appear on the trending video page consecutively based on the videos likes, dislikes, comment count, and views on the first day that it appears on the trending page. The second model predicts the number of dislikes that a video can be expected to have based on the video’s views, likes, comment count, and days trending. The reasoning for generating this model is due to the fact that YouTube no longer allows the public to see a video’s dislike counts. For these models to be deemed accurate they would have an R squared value of at least 0.7. The prediction models both had R squared values below 0.7 with the days trending model having an R squared value of 0.072. As a result of this it cannot be said that the model is accurate. A reason for the lack of accuracy could be because of limitations of the data set as well as the unpredictable nature of the culture of online video sharing. The dislikes prediction model was much more accurate than the days trending model however its R squared value was still too low for it to be considered reliably accurate. The inaccuracy of this model could also be due to the limited dataset.

**5 LIMITATIONS**

The data set that we worked with was limited in a variety of ways. There are several pieces of information that could improve the accuracy of our models such as: video length, video shares to other websites such as social media, number of subscribers that the channel publishing the trending video has, and the rank on the trending page that the video reached.

We have over 300,000 rows of data after cleaning and combining all the data from each country into one data frame. We encountered the insufficient memory issue since we did our project on Google colab and Datalore, a collaborative platform similar to Google colab.

Another limitation is the titles become unreadable if the character of the country is not in the English alphabet, so we were not able to do the sentiment analysis for those countries which was one of our initial goals. The last limitation is we have a limited dataset which is from 2017 to 2018. It would be better if we have a more recent dataset so the prediction can be more accurate to our real world nowadays.

**6 CONCLUSION and FUTURE WORK**

Throughout this paper we have examined a variety of factors and attributes that contribute to a YouTube video’s trending status. It has become clear that a majority of trending videos are from official corporate accounts. There is no one way to ensure a video will become trending or to cause a video to remain trending. A creator on YouTube may increase success by publishing the video on a weekday and making content that can fall into the Entertainment category. However, YouTube trending video performance is a complex topic with many influencing factors though there are commonalities between the most successful videos present.

Future work in this area would benefit from a larger data source. We could move forward by obtaining the data from more recent years of youTube trending videos to compare the accuracy of our model. Additionally, we could compare how the trends in category popularity of trending videos change over time. Furthermore data from additional countries could be collected if that data is available.

**7 REFERENCES**

[1] Swati Gayakwad, Rajas Patankar, Dashrath Mane (2020) Analysis on YouTube Trending Videos

https://www.irjet.net/archives/V7/i8/IRJET-V7I8732.pdf

[2] Renjie Zhou, Samamon Khemmarat‡, Lixin Gao (2010) The Impact of YouTube Recommendation System on Video Views.

https://doi.org/10.1145/1879141.1879193

[3] Anders Brodersen, Salvatore Scellato, Mirjam Wattenhofer (2012) YouTube Around the World: Geographic Popularity of Videos.

https://csci572.com/papers/YouTubeVideos.pdf

[4] Aakash Ashok Niture (2021) Predictive analysis of YouTube trending videos using Machine Learning.

https://esource.dbs.ie/handle/10788/4260

[5] Dana Adriana Lups, a-Tătaru and Radu Lixăndroiu (2022) YouTube Channels, Subscribers, Uploads and Views: A Multidimensional Analysis of the First 1700 Channels from July 2022.

https://www.mdpi.com/2071-1050/14/20/13112

[6] EDA on YouTube

https://www.kaggle.com/code/sathvisiva/eda-on-youtube

[7] YouTube likes prediction

https://www.kaggle.com/code/hetulmehta/youtube-likes-prediction/

[8] Trending YouTube Video Statistics

https://www.kaggle.com/datasets/datasnaek/youtube-new