The Effects of Like, Dislike, Number of Words in Titles, and Number of Tags on Comment Participation of YouTube Trending Videos

Chuheng(Kevin) Yu
ECON 4590 Section 01
(Dated: April 29, 2022)

Since YouTube was invented on February 14, 2005, It has become the largest video sharing platform to help creators to post video content covering various topics and created a platform for viewers to keep on track of what is happening around the world. In this paper, I analyze YouTube trending video data from ten different geographic locations: Russia, Mexico, South Korea, Japan, India, USA, Great Britain, Germany, Canada, and France from December 1st, 2017 to May 31st, 2018. Resulting in a total of 414,858 observations from the six months. The dataset includes Trending Date, Category ID, Number of Views, Number of Likes, Number of Dislikes, Number of comments, Number of words in the title, and Number of Tags. I present my findings by using explanatory data analysis among countries and categories to get data insights through data visualizations like a bar graph, dot chart, and distribution graph, and find relationships between them to understand viewers' online behavior. Moreover, I have performed OLS regression and Fixed Effect regression to measure the effects through econometrics models. Findings from this study include: More views do not lead to more comments; A viewer is more likely to comment if they like or dislike a video; Length of the title and number of tags does not directly affect the comment number. These results could provide insights for creators to maximum content exposure, and provide an overview for the viewers of how the trending section on YouTube works.

I. INTRODUCTION

YouTube was launched on Valentine's day in 2005 [6]. After seventeen years of growth, it is now the biggest video-sharing platform globally. There are 500 hours of video uploaded per minute. That is 30,000 hours of video every hour, and 720,000 hours of video being uploaded every day [8]. In total, the site gets 14.3 billion visits per month and 1.7 billion unique monthly visitors [7]. The number of views is more than on Facebook, Wikipedia, Amazon, Instagram, and only Google. YouTube is especially popular in the U.S. which 62% of American users access it daily, 92% of users claimed to use the platform weekly and 98% every month. On average, each visitors spend an average of ninety minutes a day on the platform. In 2020, 22% of YouTube users are accessing the site via mobile and tablet devices. However, the YouTube trending section did not launch the same as the website. The trending section launched in May 2005 [2]. There are currently fifty videos on the chart, and it is being refreshed every fifteen minutes. Although YouTube has an extraordinary algorithm for video suggestions, the trending chart is not personalized. No matter if you are watching it from California or Troy, it remains the same for each country.

Viewers would always be interested in how these videos raise to the chart while they were browsing. According to YouTube's explanation, there are a few qualifications [1] and factors. First, content should be appealing to a wide range of viewers within a country. Secondly, the title and video content should not contain any misleading, clickbaity, or fraud messages. Third, the chart should contain diverse creators and video content. Last but not least, the chart should be surprising and novel at the same time. There are five factors to determine if one video should be on the chart: the number of views; the "temperature" of a video(How quickly the video is generating views); where are the views coming from? (outside or inside of the official website); the age of the video; how it compared to another video in the same channel. The last factor is mainly for promoting artists or video content. If a viewer clicks the trending section, "Artist on the Rise" would always appear at the top of the screen. The quote from the YouTube Support website perfectly summarized the purpose and the vision of the trending chart, "We combine these signals to produce a list of videos that showcases what's happening on YouTube.".

Before bringing the data into models, I used exploratory data analysis (EDA) to highlight characteristics. It is a useful procedure to understand the data set fully and to identify the relationships between attributes through visualizations and distributions. Moreover, EDA is useful to identify trends and test fundamental assumptions. Visualization methods like a bar graph, dot graph, and distribution graph allow the data to reveal its structure if it is utilized properly. The EDA analysis is not limited to graphical analysis but also includes the following purposes [5]:

- get a deeper insight into the present dataset;
- discover optimal factor settings;
- develop penurious predictive models;
- test fundamental assumptions;
- find outliers and anomalies in the dataset;
- extract useful variables and information;
- reveal hidden structures.

Later in this paper, I will use pooled OLS regression and Fixed Effect regression to measure the effects of independent variables. A multiple regression model is an extension of the one dependent variable matching one independent variable regression model which has multiple regressors with coefficients, an intercept, and an error term. Compared to a simple regressor model, researchers could compare the effects of various factors and determine the relationship between them. As for the Fixed Effect model, the alpha i that represent a fixed amount for each individual would avoid omitted variable bias when the independent variable is correlated with observable factors. Therefore, the result should be consistent and unbiased when I implement these two models with the panel data set.

This research will be conducted on RStudio and focus on analyzing data trends and viewers' comment participation behavior. With packages like *lmtest*, *plm*, and *dplyr*, hidden patterns would be discovered from all geographic locations and categories. There are also a few challenges involving interaction terms with pooled OLS regression, which will be discussed in the later section of the paper.

II. LITERATURE REVIEWS

Even though YouTube is such a huge platform with plenty of data, there are a few studies using econometric models that focus on commenting behavior. Although viewers' comment behavior is affected by external events globally, a variety of views and video content represent characteristics and trends worldwide. And it is complex and difficult to have a 100% accuracy to generate predictions from these attributes.

A paper was published by Ilana Dubovi and Iris Tabak in 2018 [3] studying the emotional and cognitive engagement with Science content on YouTube. This paper explained how viewers' comment behavior with videos in the Science and Education category. In the data collection process, they used YouTube API which is commonly used to extract information on likes, dislikes, number of comments, comments, and descriptions. This dataset contains the number of views, likes, dislikes, comments, and a random sample of 1000 comments per video from March 2019 to June 2019 with 200 trending videos per day. They use two methods to analyze the data: one-way analysis of variance(ANOVA) and the Syuzhet package. ANOVA provides a statistical test of whether two or more population means are equal, and therefore generalizes the t-test beyond two means. In other words, the ANOVA is used to test the difference between two or more means. Syuzhet, an R package, enables sentiment polarity evaluation and emotional categorization into two segments, positive and negative. These two segments, it divides into eight types: anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. Before implementing these methods, the authors used explanatory data analysis to have a brief overview of the dataset. Within the dataset, videos in the music category have a higher mean number of views and comments which implies that videos in the music category are more likely to get viewers to engage as in figure 1. Also, in the Howto & Style category, viewers are more likely to discuss in the comment section compared to their average of views. This shows that after viewers check out a video in the Howto & Style category, they would more likely to go back and share their experience and provide feedback. From the comment sentiment analysis of 1000 random comments per video, the author found that trust has the most expressions followed by anticipation and fear. There is no specific pattern between positive and negative emotions. Therefore, one conclusion is as more comments, the sentiment of the comment section is more neutral. Although Science & Education category only takes 3.4% of the total 14,700 observations, which Entertainment category is 30%, the Music category is 14.5%, and the Sports category is 11.4%, it has a higher average of post-video comments on average with mixed emotion expressions. The paper provides me with a global trend among categories, and set a starting point for my study: viewers are more likely to comment when they want to express their emotions about the topic and the content of the video.

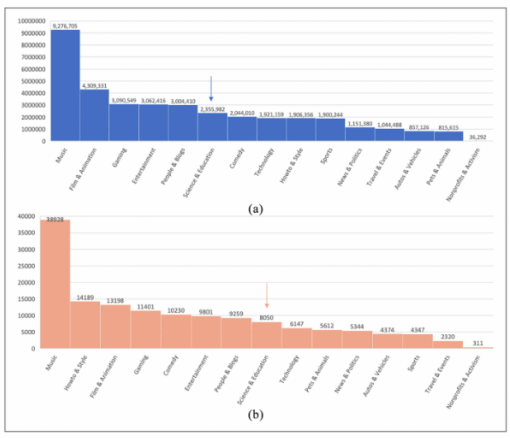

FIG. 1. (a)mean number of views per category and (b) mean number of post-video comments per category.

The second paper I would like to discuss is published in 2018 about socio-computational analysis of trending YouTube videos [9]. In this research, they studied the top 200 YouTube videos trending daily for 40 days from September 2017 to October 2017 in the United States of America and the Great Britain region. The dataset's size is approximately 8000 videos for each region. This research uses the same attributes *title of the video, URL of the video, video ID, number of comments, number of views, number of likes,* and *number of dislikes.* Later on, they added the *description of the video, date the channel was created* and *the number of subscribers of the channel.* The authors extracted associations with other social media platforms that are attached through web links by

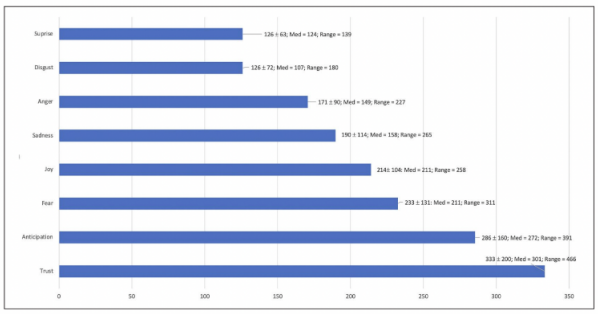

FIG. 2. Sentiment analysis: Amount of emotional expressions per video (Mean $\pm SD$; $median$; $10 to 90 percentile range$.

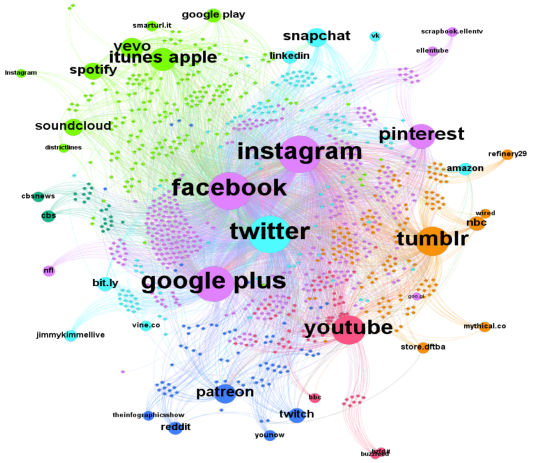

FIG. 3. Social media map of the trending videos in USA

using Web Content Extractor to study cross-media integration. The network map in Figure 3 that helps the understanding of the relationships between social media platforms shows the cluster of the most commonly used social media websites and how these connect. Social media related to the Music category (green colored nodes) clustered independently. This implies that videos in the Music category mostly are shared to completely different media sites where creators and artists could earn direct income from it.

To examine users' engagement behavior on YouTube, the researchers did correlation analysis among *number of views, number of likes, number of comments, sentiments from comments* and *sentiments from video description*. Within this study, they observed that if a viewer likes or views a video then he/she is more likely to comment on that video due to a high correlation between the number of comments, number of views, and number of likes. Additionally, they observed that as more comments in a video, the more neutral the sentiments are in the comment section. This finding also explains the result from Ilana's paper [3] which the emotional categorization analysis shows the sentiments in the comment section of Science & Educational video content are mixed between positive and negative emotions. Also, a similar trend is observed between description sentiment and the number of likes: the more likes in a video, the more neutral the description sentiment is. These findings provide a tremendous guide for creators to a maximum number of likes, the number of views, and the number of comments by ingeniously using description and comment strategies with attractive content.

III. METHODOLOGY

A. Dataset Overview

In this study, the dataset is obtained through Kaggle created by Mitchell J[4] in 2019. The dataset extracted data from 12/01/2017 to 05/31/2018 with 200 videos per day. The video data are from ten different geographic locations: Russia, Mexico, South Korea, Japan, India, the United States of America, Great Britain, Germany, Canada, and France. There are in total 414,858 observations from 32 categories. Within the raw dataset, there are 16 attributes: video's ID, trending date of the video, title of the video, channel title, category id, the publish time of the video, a list of tags, number of views, number of likes, number of dislikes, number of comments, URL link of the video, comments disabled, ratings disabled, video error or removed, and the description of the video where comments disabled, ratings disabled, video error or removed are binary variables.

B. Exploratory Data Analysis

Before editing the database and manipulating it into a panel dataset, I apply techniques such as histogram and dot graph to present raw data mean, plots, and others for presenting statistics and trends. The reason I choose this method is EDA could analyze and explain the dataset from a quantitative way also avoiding biased information and outliers to support my regression results. In this step, I try to find out as much trends and correlations as possible which might not be shown in my regression models. The followings are the main interesting findings and relationships between *the number of views, number of likes, the number of dislikes, the number of comments, category ID, country name, length of words in the title*, and *number of tags* in my dataset. First, I conducted a simple histogram which has category ID as independent variable and the number of views as the dependent variable as in the Figure 4. As similar findings from previous literature studies, Music category (category ID:10) has the most views following by Entertainment (category ID:24) and Film & Animation category (category ID:1). Although the second and third most viewed category might be affected by external events, Music category

would always hold the most viewed category. This also implies on YouTube, Music category has the most audience size and attention. This finding also provides an suggestions for entry creators that if they want to gain attention and streaming flow quickly, posting music related content should be the first option.

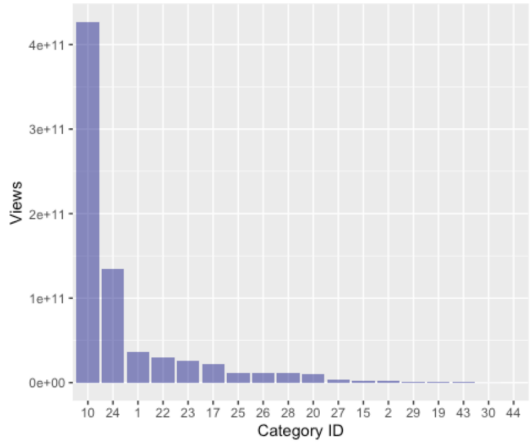

FIG. 4. The number of views by categories

Next, I studied data by country and explore how people from different countries participate on the YouTube platform. Before starting to make further calculations, I first aggregated all the quantitative attributes by country. For this observation, I calculated the likes percentage, dislikes percentage, and comment percentage by dividing the number of likes, dislikes, and comments by the total number of views. Then generated bar graphs to represent the findings. As is shown in Figure 5, Russia has the highest number of likes per view followed by Mexico and France, while Great Britain, India, and South Korea have the lowest percentage. The dislikes per view by country are more skewed to the right 6. Russia has the highest dislikes percentage among the ten countries, even higher than Germany, Mexico, and France combined. Therefore, Russian viewers are more polarized through their online actions while having nearly the same number of views as the other nine countries. Last but not least, I also did a comments percentage per country to test my assumption about Russian viewers. Once again, Russia is leading the chart with 0.15% more than Mexico. Consequently, viewers from Russia are more active by clicking the like and dislike button and posting comments to share their thoughts. However, this could not be certain because due to the internet closure in countries like China because people use VPNs to view website content they are interested in. So a further study focusing on Russia through a longer period with a comparison with other countries is suggested to figure out the true reason behind it.

For qualitative attributes such as the title of the video and description of the video, I added *tag_count* and *title_count* representing the number of tags a video has and the length of the video's title. With these two new

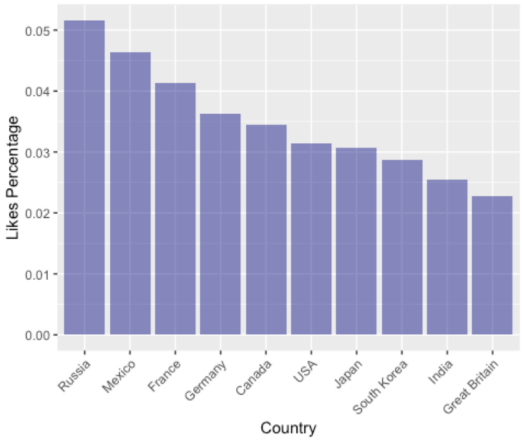

FIG. 5. Likes per view by country

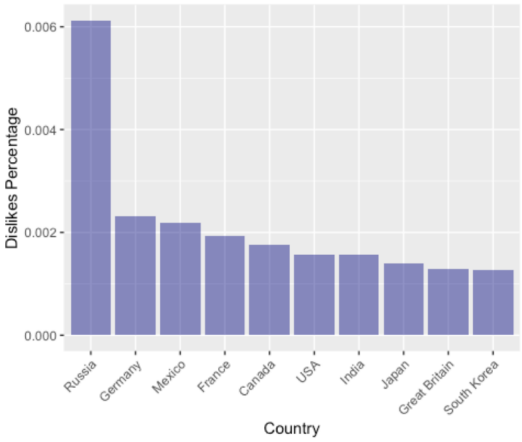

FIG. 6. Dislikes per view by country

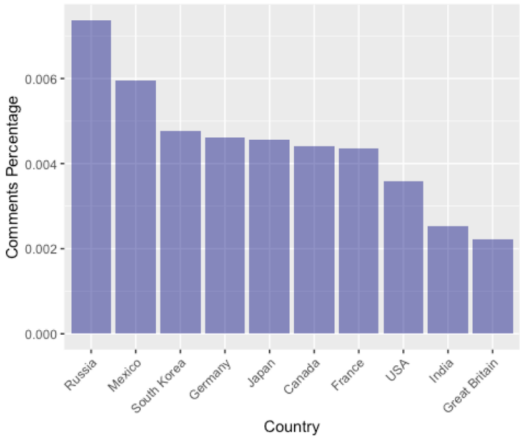

FIG. 7. Comments per view by country

variables, I conducted histograms containing views and the number of comments as dependent variables for each of the new variables contain. From Figure 8 and Figure 9, all four of the graphs are right-skewed. Despite the outliers, we can see videos has the highest views with

10 to 30 tags and 5 to 10 words in the title. A similar pattern is found in Figure 9 where videos have the maximum number of comments when it has 10 to 30 tags and 4 to 10 words in the title. Video creators and artists in this period have found the right range for the number of tags and length of the title to maximize their number of views and viewers' comment participation. In general, a video with 20 to 30 tags and a title length from 6 to 10 words will have a higher chance to gain maximum views and many comments. But these attributes are not the determining factors since there are outliers in all four graphs. As supporting factors, creators could use these findings as suggestions to avoid unnecessary viewer loss. Therefore, even if I will include the number of tags and the number of words in the title, I speculate these two attributes would not have a significant effect on the number of comments.

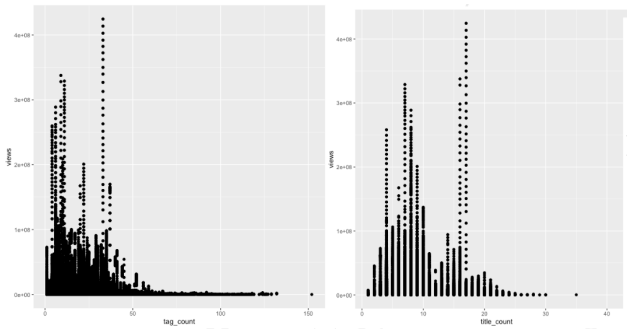

FIG. 8. Number of tags and length of the video's title versus the number of views

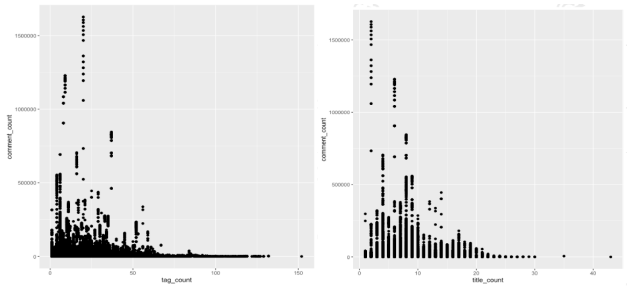

FIG. 9. Number of tags and length of the video's title versus the number of comments

C. Data cleaning and pre processing

This is the first step before I bring the data into any regression models. Because the raw data set is not in a time series or panel format, I first aggregated every dataset from different countries into one. Then, I dropped videos which are removed or had their comment section disabled with binary variable *video_error_or_removed* and *comments_disabled* and dropped them after filtering out the data. Then dropped rows with NA and Null values to avoid any errors in further tests. By aggregating the dataset through trending date and the category it is in, a panel dataset is formed.

D. Regression Analysis

In this section, I brought the panel dataset into Pooled OLS regression and Fixed Effect regression model. In this research, I was mainly focusing on how the number of comments is affected by the number of likes, the number of dislikes, the number of views, the length of the video's title, and the number of tags. To represent multiple groups in a single regression equation, I added dummy variables for each of the categories besides the first category. The regression model can be expressed as the following:

$$comments\_count_i = \beta 0 + \beta 1 Views_i + \beta 2 Likes_i + \beta 3 Dislikes_i + \beta 4 Title\_count_i + \beta 5 Tag\_count_i + \varepsilon_i$$

Where i represent the category ID

In this model that is shown in Figure 10, the number of views has relatively no effect on the number of comments, and its coefficient is significant. Both numbers of likes and dislikes have a significant coefficient and have a positive impact on the number of comments. While the length of the title and the number of tags have a high coefficient, none of them are significant. From this regression model, I found there are a couple of problems from this model. First, because of the large scale of *comments_count*, *views_count*, *likes_count*, and *dislikes_count*, the coefficients are relatively small but significant, and the coefficients for *title_count* and *tag_count* are large and insignificant due to the small range comparing to other attributes. *title_count* is not significant at 1% significance level, and *tag_count* is not significant at all. The coefficients of both of *title_count* and *tag_count* are larger than 1. Therefore, I modified my model as the following:

$$log(comments\_count_i) = \beta 0 + log(\beta 1 Views_i) + log(\beta 2 Likes_i) + log(\beta 3 Dislikes_i) + \beta 4 Title\_count_i^2 + \beta 5 Tag\_count_i^2 + \varepsilon_i$$

Where i represent the category ID

The reason I modified my model in this way is that using log and squared values could transform skewed variables into a more normalized dataset while maintaining the characteristics and trends of all the attributes. The improved version of the model has a more reasonable result as is shown in Figure 11. In the modified model, the coefficient of the log of the number of views has increased while not being significant. While the log of the number of likes and the number of dislikes remain significant, the impact of the number of likes decreases, and

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2.1649e+05	3.2396e+04	6.6828	2.746e-11	***
views	-1.2041e-03	2.2750e-04	-5.2928	1.284e-07	***
likes	1.2917e-01	1.3482e-02	9.5811	< 2.2e-16	***
dislikes	5.1480e-01	3.6546e-02	14.0862	< 2.2e-16	***
title_count	-1.2185e+02	4.7543e+01	-2.5629	0.0104253	*
tag_count	-1.7930e+01	2.6399e+01	-0.6792	0.4970540	
category_id_2	-1.7902e+05	2.5519e+04	-7.0151	2.781e-12	***
category_id_10	-9.2687e+05	1.8975e+05	-4.8846	1.086e-06	***
category_id_15	-1.7344e+05	2.5301e+04	-6.8550	8.490e-12	***
category_id_17	7.2064e+03	1.5744e+04	0.4577	0.6471836	
category_id_19	-1.9688e+05	2.9338e+04	-6.7108	2.273e-11	***
category_id_20	-9.2276e+04	2.1113e+04	-4.3705	1.278e-05	***
category_id_22	1.9927e+05	4.4168e+04	4.5115	6.664e-06	***
category_id_23	-1.4230e+05	4.0099e+04	-3.5487	0.0003925	***
category_id_24	8.8856e+05	1.4571e+05	6.0983	1.197e-09	***
category_id_25	1.9551e+05	3.2316e+04	6.0500	1.613e-09	***
category_id_26	-1.6039e+04	1.2979e+04	-1.2357	0.2166640	
category_id_27	-1.6099e+05	2.1021e+04	-7.6589	2.455e-14	***
category_id_28	-6.3788e+04	2.6441e+04	-2.4125	0.0159001	*
category_id_29	-1.7041e+05	3.2866e+04	-5.1850	2.291e-07	***
category_id_30	-2.1331e+05	3.1631e+04	-6.7435	1.821e-11	***
category_id_43	-2.0901e+05	3.0965e+04	-6.7500	1.742e-11	***
category_id_44	-2.1588e+05	3.2322e+04	-6.6791	2.815e-11	***

FIG. 10. Pooled OLS Regression Model

the impact of the number of dislikes increases. Therefore, when viewers dislike a video, they are more likely to express their emotions in the comment section. Also, views do not have a significant coefficient throughout all three models. So the common sense that "More views in the video, the more comments there are." is proved wrong. Moreover, *title_count* and *tag_count* became insignificance with a large decrease with its coefficients. Because I aggregated the dataset by category, most of the categories from the first OLS model are significant which does not match our findings in previous explanatory data analysis graphs. From the modified model, Auto & Vehicles, Gaming, News & Politics, Howto & Style, Science & Technology, Movies, and Shows are the categories that have a significant coefficient. Among all the significant categories, Gaming, News & Politics, Howto & Style, and Science & Technology have a positive coefficient which means that if a video is in one of these categories, it has more comments in general, despite the country and creator. Because of the differentiation between categories, creating interaction terms between views and category ID should be helpful to identify which category has a more positive or negative effect on comment participation. Therefore, I used Fixed Effect regression model and created interaction term with *log(views* and *category_id*. As is shown in Figure 12, the result provides more information. Of all the interaction terms between the number of views and the category ID, there are only a few that stands out and be significant. Categories like Auto & Vehicles, Travel & Events, Gaming, Comedy, Howto & Style, Science & Technology, and Nonprofits & Activism are extremely significant. And besides Gaming, Comedy, and Howto & Style, all other significant categories has a positive coefficient which means when views increases in these categories, it would lead to an increase in comment number.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2.9020e-01	2.1309e-01	1.3618	0.173348	
log(views)	-2.6437e-02	1.8490e-02	-1.4298	0.152868	
log(likes)	6.7595e-01	2.0997e-02	32.1926	< 2.2e-16	***
log(dislikes)	2.3753e-01	1.0883e-02	21.8259	< 2.2e-16	***
title_count	1.4073e-05	3.3458e-05	0.4206	0.674066	
tag_count	2.7949e-05	1.6039e-05	1.7426	0.081505	.
category_id_2	-1.6696e-01	3.9206e-02	-4.2586	2.115e-05	***
category_id_10	-1.0990e-01	4.2463e-02	-2.5881	0.009694	**
category_id_15	5.4681e-02	3.8032e-02	1.4378	0.150592	
category_id_17	-4.4890e-02	2.0719e-02	-2.1666	0.030340	*
category_id_19	-5.4676e-02	4.7317e-02	-1.1555	0.247956	
category_id_20	2.1918e-01	3.0463e-02	7.1947	7.730e-13	***
category_id_22	2.2111e-02	3.7881e-02	0.5837	0.559460	
category_id_23	-3.5454e-02	2.5334e-02	-1.3995	0.161769	
category_id_24	-4.9122e-02	9.1088e-02	-0.5393	0.589731	
category_id_25	4.6489e-01	3.7272e-02	12.4731	< 2.2e-16	***
category_id_26	1.5461e-01	3.0280e-02	5.1061	3.477e-07	***
category_id_27	-6.4385e-02	2.9595e-02	-2.1755	0.029662	*
category_id_28	2.1530e-01	3.6960e-02	5.8253	6.254e-09	***
category_id_29	-9.8562e-02	5.8603e-02	-1.6819	0.092689	.
category_id_30	-9.7077e-01	9.7318e-02	-9.9752	< 2.2e-16	***
category_id_43	-4.0319e-01	6.1511e-02	-6.5547	6.455e-11	***

FIG. 11. Pooled OLS Regression Model with Modified Variables

	Estimate	Std. Error	t value	Pr(>|t|)	
log(views)	-1.0518e-01	4.4995e-02	-2.3376	0.019469	*
log(likes)	6.9533e-01	5.1700e-02	13.4495	< 2.2e-16	***
log(dislikes)	2.3735e-01	2.6107e-02	9.0915	< 2.2e-16	***
title_count	2.3707e-07	7.6246e-05	0.0031	0.997519	
tag_count	4.0471e-05	5.4873e-05	0.7375	0.460852	
log(views):category_id_2	1.6185e-01	2.5176e-02	6.4288	1.474e-10	***
log(views):category_id_10	9.8311e-02	3.1153e-02	3.1557	0.001616	**
log(views):category_id_15	2.6914e-02	9.2299e-03	2.9160	0.003570	**
log(views):category_id_17	2.2066e-02	2.2373e-02	0.9863	0.324068	
log(views):category_id_19	1.6888e-01	2.3328e-02	7.2394	5.598e-13	***
log(views):category_id_20	-1.6516e-01	1.5160e-02	-10.8947	< 2.2e-16	***
log(views):category_id_22	8.2244e-02	2.5393e-02	3.2388	0.001212	**
log(views):category_id_23	-2.0817e-01	1.2423e-02	-16.7570	< 2.2e-16	***
log(views):category_id_24	3.7574e-02	4.8871e-02	0.7688	0.442047	
log(views):category_id_25	-2.6508e-03	9.6976e-03	-0.2733	0.784605	
log(views):category_id_26	-1.8782e-01	2.4032e-02	-7.8156	7.333e-15	***
log(views):category_id_27	6.2800e-03	1.6879e-02	0.3721	0.709877	
log(views):category_id_28	1.7489e-01	1.5456e-02	11.3157	< 2.2e-16	***
log(views):category_id_29	1.8842e-01	2.1029e-02	8.9598	< 2.2e-16	***
log(views):category_id_30	2.2154e-02	1.4955e-02	1.4814	0.138586	
log(views):category_id_43	-4.5000e-02	1.3966e-02	-3.2222	0.001285	**

FIG. 12. Fixed Effect regression model with interaction terms between the number of views and category ID

IV. CONCLUSION AND FUTURE WORKS

The analysis is performed on YouTube data collected over 205 days which contains 414,858 videos. At first, I used bar graphs to generate an overview of my dataset. At first, we figured out the overall trends for categories and countries with the number of views, likes, dislikes, and comments. From the analysis, we can see that category that has the most views is Music, and following by Entertainment, Film & Animation. And from my studies between geographic locations and viewers' behaviors, the country that has the most like per view is Russia, followed by Mexico and France. The top 3 countries with

the most dislike percentage are Russia, Germany, and Mexico. And the country that has the most comments per view is Russia followed by Mexico and South Korea. From the last three graphs, Russia has the most polarized viewers which are leading all three charts. Therefore, future studies mainly focus on Russian comment sentiments with packages like Syuzhet in a longer period could find out the reason behind these behavior trends. However, due to the increase of people using the VPN to protect their privacy and to access location-based web content, the country data could not be fully trusted. Some potential reason behind it would be viewer from China uses a Russian server VPN to access YouTube content. Later on, I tested the relationship between the length of videos' titles, the number of tags, and the number of views and comments. When a video has 20 to 30 tags and a title length from 6 to 10 words, it is most likely to have the maximum views and comment participation. A future study that focuses on a certain category or particular creators and observes the relationship between these attributes through a 2 to 5 years period should bring a more clear vision of how title length and tag number affect a video's performance. From the regression model, the coefficients and significance level proved a lot of our assumptions from previous sections, but it gives a more specific relationship between categories. I observed that the effect of views, length of the title, and tags are not significant.

Because the dataset is from 2017, one concern I had is it might not represent the current situation. YouTube also made some changes to its platform which might affect this study. First, they integrated the "short video" form with their previous streaming service. I wanted to focus on regular-length video comment participation in this study mainly because shorts do not have a dislike function. Secondly, YouTube removed the dislike counter to the public at the beginning of 2022. This directly affects how people judge a video's quality which might lead to an increase in the comments. Therefore, a newer dataset is needed if researchers want to figure out how the changes would affect people's online behavior. Efficient use of regression models and EDA can testify to the assumptions and improve the accuracy of further models by identifying errors, relationships, and unnecessary data.

[1] YouTube Help Center. Trending on youtube, 2022.

[2] Megan Rose Dickey. The 22 key turning points in the history of youtube, 2013.

[3] Ilana Dubovi and Iris Tabak. Interactions between emotional and cognitive engagement with science on youtube. *Public Understanding of Science*, 30(6):759–776, 2021. PMID: 33546572.

[4] Mitchell J. Trending youtube video statistics and comments, 2019.

[5] Sana Khanam, Safdar Tanweer, and Syed Sibtain Khalid. Youtube Trending Videos: Boosting Machine Learning Results Using Exploratory Data Analysis. *The Computer Journal*, 10 2021. bxab142.

[6] Christopher McFadden. Youtube's history and its impact on the internet.

[7] Stacey McLachlan. 23 youtube stats that matter to marketers in 2022, 2022.

[8] Maryam Mohsin. 10 youtube stats every marketer should know in 2021 [infographic].

[9] Samer Al-Khateeb Kiran Kumar Bandeli Nitin Agarwal Muhammad Nihal Hussain, Serpil Tokdemir. Understanding digital ethnography: Socio-computational analysis of trending youtube videos, 2018.