Trend analysis and prediction of YouTube Videos using Machine Learning Techniques.

DHRUV PATEL[1], DHRUVI MODI[1], NIMA PATEL[1]
1. BTech - CSE, Indus Institute of Technology, Ahmedabad

I. ABSTRACT

Over the past few years, India has seen tremendous growth in internet penetration. In 2012, internet penetration was around 12.6% which has increased to 50% in 2020. The major source of internet consumption in India is through mobiles and according to a report, there are 356 million mobile users who engage in video content. The most preferred platform for video content is YouTube because it has tons of free videos from every category which draws viewers' attention through its smooth user interface which has led to its popularity among all generations. The youtube trending page shows videos that attract a large number of viewers and which have a high like to dislike ratio. This research paper analyses the engagement of viewers with various content categories and shows the prediction of the number of likes and views. Youtube trending data is used as a dataset for this research. The research revealed that in the majority country music video category is the most engaged category among other categories and comedy is the second most engaged category. Whereas, News & Politics has the lowest engagement. Various models were applied to find the best results in which KNN best fits our model with the highest accuracy of about 61.4%.

II. KEYWORDS

Machine Learning, Data visualization, YouTube Trend Analysis, Best Video Category

III. INTRODUCTION

The Internet has marked its most remarkable impact on the journey of every human being. With the gradual increase of internet users, every domain has seen a sprouting growth in them. As we speak of categories that have expanded with the increase in internet penetration, the entertainment industry has become catchy to almost all age groups. Earlier there were limited means of entertainment encountered by people as only movies and few tv shows were available. Subsequently, the situation is different now as there are several platforms showcasing entertainment, resulting in the popularity of social media platforms such as Facebook, Instagram, Twitter, and video streaming platforms like YouTube along with OTT platforms like Netflix, Amazon prime, etc. YouTube is so far one of the easiest video-sharing platforms that allow users to upload their videos along with visiting videos posted by other creators. It offers content creators a great platform to share knowledge, ideas, and interesting information with their viewers and can hook their attention. YouTube has a variety of genres ranging from Comedy videos, Product Reviews, DIY/Tutorials, Commentary Videos, Live Streaming, Vlogging, Top List, Reaction Videos, and Q&A-type videos to sketch videos, short films, and even some video series. Due to its free videos and diverse content, YouTube is becoming one of the most viewed video-sharing platforms in India with the most YouTube users in 2021, estimated at 225 million. Youtube has several features in addition to easy video sharing which includes a strong recommendation algorithm based on users' frequently viewed content, youtube shorts that are 10-60 seconds videos, live streaming, subscriber notifications, and the trending page that will show the latest feeds daily.

IV. Related Work

In recent times, various works have been done related to youtube video trends. One of them is to predict the emerging trend on social media platforms[i]. This paper works on the Growth-based Popularity Predictor (GPP) model for predicting and ranking the web-contents. It uses data from Movielens, Facebook-wall-post, and Digg-like platforms to conduct studies and create a model to identify future growth.

Another work in a similar field is conducted by Amar Krishna, polarity trend analysis of public sentiment on youtube[ii]. In this paper, he has performed sentiment analysis using comments on the videos. Data for the research were accumulated from youtube and used various machine learning techniques to shape the research. It demonstrates that an analysis of the sentiments to identify their trends, seasonality, and forecasts can provide a clear picture of the influence of real-world events on user sentiments. Results show that the trends in users' sentiments are well correlated to the real-world events associated with the respective keywords.

Moreover, the technique of Big data is also used in Program Popularity Prediction in Broadcast TV Industries[iii]. The research was conducted by CHENGANG ZHU, GUANG CHENG1, AND KUN WANG. They are senior members of IEEE. A few years back TV industries were more famous and had the greatest attention of the public which is now slowly overtaken by youtube. This research paper provided very useful data to formulate purchasing decisions and also help to formulate some critical financial spending decisions. This paper has applied a distance-based K-medoids algorithm to group programs' popularity evolution into four trends. Then, four trend-specific prediction models are built separately using random forest regression.

V. BACKGROUND THEORY AND FEATURE ANALYSIS

There is a good deal of elements that go around a video that could catch the eye of the intended audience. These elements include:

Title:- It summarises the content of the video and includes keywords that help users to find those videos.

Description:- A YouTube video description is the text below each of your videos. These little text nuggets help viewers find your content and decide whether to watch it. YouTubers can also add some hashtags to trend the video for some specific video audiences.

Likes and Dislikes:- Youtube provides a like button feature to show the creator that you enjoyed their video and appreciated their work. The like and dislike count shows the stats of the number of people who liked or disliked the video.

Comment Section:- In this section users can put their thoughts about videos and share their feedback. The comment section creates engagement between video viewers and video creators. The Youtube algorithm uses these engagement stats to promote and recommend videos to other users.

Views count:- It shows the number of views on a particular video.

Video thumbnails:- A video thumbnail is like a still image that acts as the preview image for the video.

Table V.I: Dataset's features and their description

No	Feature Name	Data Type	Description
1	video_id	Alphanumeric string	Unique identifier for a video
2	title	String	Title of the video
3	publishedAt	Timestamp	Date and time of uploading
4	channelId	String	Unique identifier for a youtube channel
5	channelTitle	String	Title of the channel
6	categoryId	Integer	Id assigned to a particular genre
7	trending_date	Timestamp	Date when the video was on the trending list
8	tags	String	Keywords related to the video
9	view_count	Integer	Number of views on video
10	likes	Integer	Number of likes on the video
11	dislikes	Integer	Number of dislikes on the video
12	comment_count	Integer	Number of comments on the video
13	thumbnail_link	_URI	Link of the thumbnail image
14	comments_disabled	Boolean	Whether the comment section is disabled
15	ratings_disabled	Boolean	Whether rating of videos is disabled
16	description	String	Description about video

From the above features 4 features i.e views_count, likes, dislikes, comment_count were used for the prediction part and features like view count, likes, dislikes, comment count, comment disabled, rating disabled were used for the analysis of the data.

Table V.II Category List and Description for Youtube Videos

Category ID	Category Name
1	Film & Animation
2	Autos & Vehicles
10	Music
15	Pets & Animal
17	Sports
19	Travel & Events
20	Gaming
22	People & Blogs
23	Comedy
24	Entertainment
25	News & Politics
26	Howto & Style
27	Education
28	Science & Technology
29	Nonprofits & Activism

VI . RESEARCH METHODS

This research uses a youtube trending video dataset from May 2021 to February 2022. The research was conducted in two phases - the first phase of research uses various data visualisation techniques to find the relation between dataset features. Various graphs were created during the process and conclusions based on that were written. In the second phase of the research, machine learning models were used to predict the video category based on the features.

The first step in phase two was to collect data. Data was collected from Kaggle which is gathered by the youtube trending Video API provided by youtube.

The second step was to clean the data, fill the missing values in the dataset and find the outliers.

The third step includes applying various machine learning models on various features and finding out the best features and models to conclude this research.

Fourth Step conclusions were written based on the results acquired from applying various machine learning algorithms.

VII. Implementations and Results

This research work has gone through 2 phases. In the first phase, trending page data of various countries were analyzed based on their video's view count, video category, likes, dislikes, and comments count. As a result, we separated each category into high, medium, and low engagements in the video category.

Table VII.I : Country wise Video Engagement

Country	Engagement	Content Category
India	High	10,15,26
	Medium	12,22,23,24,20,17,19
	Low	25,26,27
Brazil	High	10,20,22,23,24,28,29
	Medium	1,5,17,25,26,27
	Low	2,19
Canada	High	10,22,23,24,29
	Medium	1,15,17,19,20,26,27,28
	Low	2,25
Germany	High	1,10,23,27,28
	Medium	17,19,20,22,24.26,29

	Low	2,25
France	High	10
	Medium	20,22,23,24,28,29
	Low	1,2,15,19,25,36,27
Japan	High	10,28,29
	Medium	17,19,20,22,23,24,25,26,27
	Low	1,2,15
Great Britain	High	10,29,22,23,24
	Medium	1,15,17,19,26,27,28
	Low	25
S.korea	High	10
	Medium	29,20
	Low	1,2,15,17,19,23,24,25,26,27,29

For phase two :

Algorithm	Absolute Error	RMSE
Linear Regression(before cleaning)	1670548.17	4399495.26
Linear Regression(after cleaning)	1683699.80	4806236.50
Linear Regression(after cleaning) - 2	1660956.5643865475	4727535.52978111
Linear	1531497.89	3630887.79

Regression(after cleaning) -3		
Polynomial Regression degree =2	1500602.06	3553338.06
Poly degree=3	1445384.35	3372055.87
Poly degree=4	1423558.85	3299634.55
Poly degree=5	1412700.00	3257163.25
Poly degree=6	1384146.38	3163764.32
Ridge Regression alpha 10	1462189.60	3530730.35
Ridge 5.5(best alpha) L2	1460946.04	3549343.423
Lasso cv	1497387.70	3529733.17

Table VI.II Results of Predicted Various Models

Algorithms	Features	Accuracy
Logistic regression	Vs category id	0.39971580443134574
KNN	Vs category id	0.6146518604283985
SVC linear	Vs category id	0.4019788
SVC rbf	Vs category id	0.416346
Decision Tree	Vs category id	0.27406
Random Forest	Vs category id	0.325

The Knn gave best results with 61.4% of accuracy Other models like Logistic regression yield 39.9% accuracy ,SVC (41.1%), random forest (32.2%) and decision tree with 27.4% accuracy.

VIII. FIGURES:

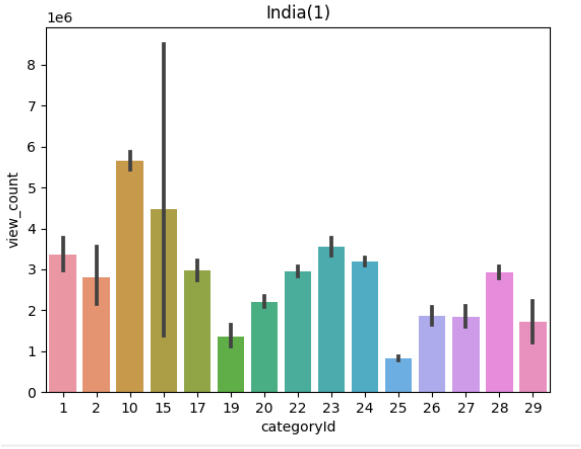

Figure 1.1: View Count vs. CategoryId

Basic conclusions:

Category ID-10 (MUSIC) has a maximum average view among all categories

Category ID-25 (News & Politics) has a minimum average view among all categories

Category ID-15 (Pets & Animals) has the highest deviation in average views

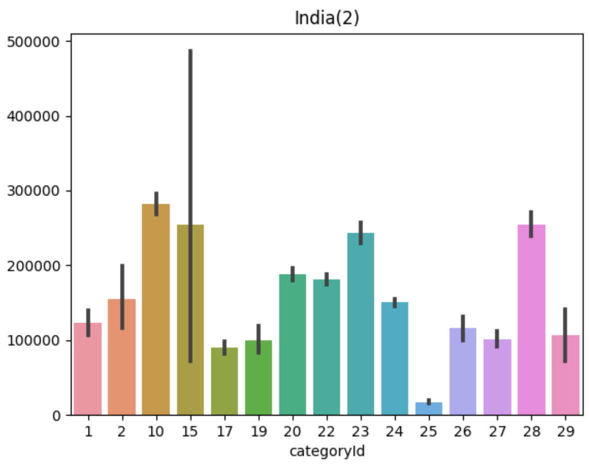

Figure 1.2 : Likes vs. CategoryId

Basic conclusions:

Category ID-10 (MUSIC) has maximum average likes among all categories

Category ID-25 (News & Politics) has a minimum average view among all categories

Category ID-15 (Pets & Animals) has the highest deviation in average views

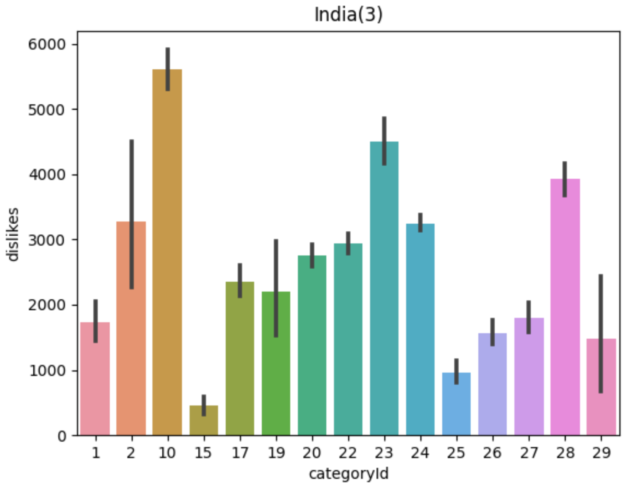

Figure 1.3: Dislikes vs. CategoryId

Basic conclusions:

Category ID-10 (MUSIC) has a maximum average dislikes among all categories
Category ID-15 (Pets & Animals) has a minimum average dislikes among all categories
Category ID-2 (Autos & Vehicles) has the highest deviation in average dislikes
Category ID-23 and 28 (Comedy and Science & Technology) has a relatively high dislike

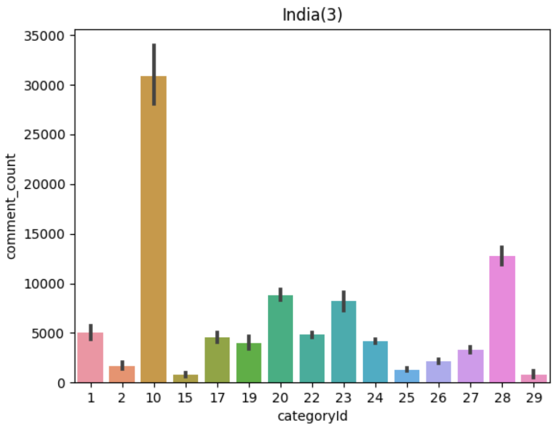

Figure 1.4 : CommentCount vs. CategoryId

Basic conclusions:

Category ID-10 (MUSIC) has a maximum average comment count among all categories.
Category ID-29 (Nonprofits & Activism) has a minimum average comment count among all categories.
Category ID-10 (music) has the highest deviation in average views

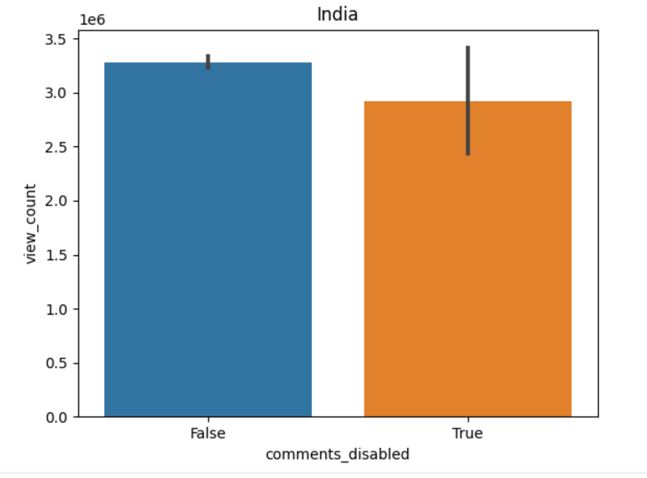

Figure 1.5: Impact of comments disabled on views

Basic conclusions:

There is not a major difference in average view count whether a comment is disabled or not, but there is a huge standard deviation in views with a comment disabled.

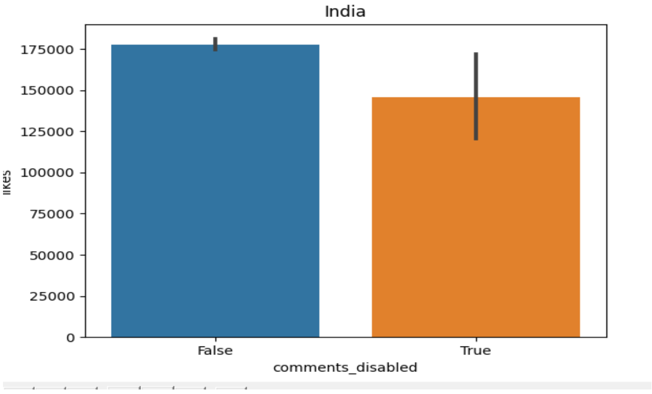

Figure 1.6: Impact of comments disabled on Likes

Basic conclusions:

There is a noticeable difference in likes after enabling the comment section and the standard deviation of disabled is also high.

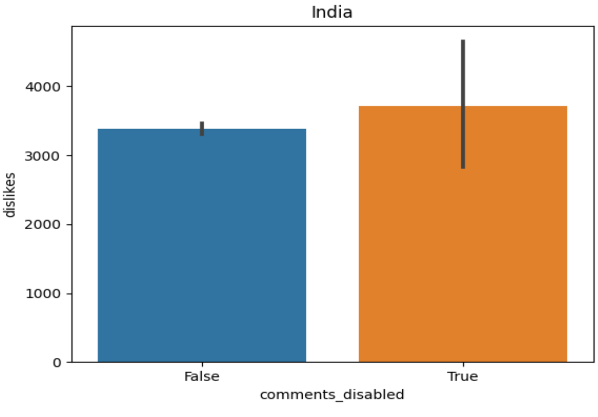

Figure 1.7: Impact of comments disabled on Dislikes

Basic conclusions:

Video's average dislike is greater when the comment section is disabled

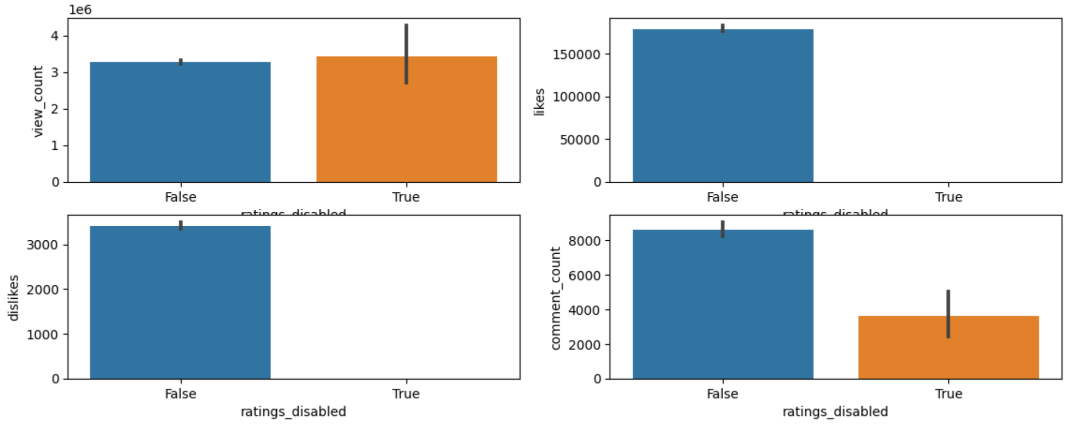

Figure 1.8: Impact of rating disabled

Basic conclusions:

There is no significant difference in view count whether the rating is disabled or not but it has a noticeable effect on comment count when rating is disabled.

COUNTRY WISE ANALYSIS:

INDIA:

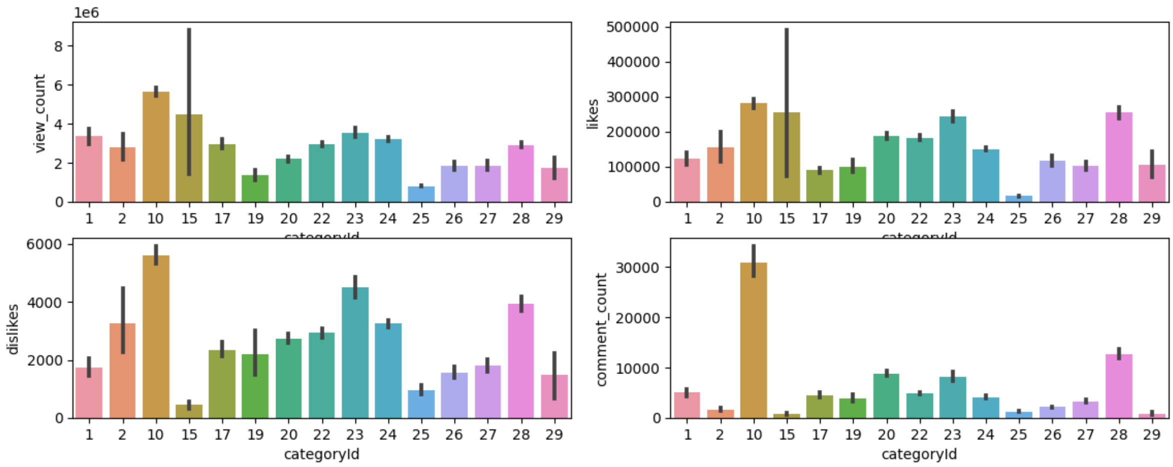

Figure 1.9: Analysis of Video engagement in India

BRAZIL:

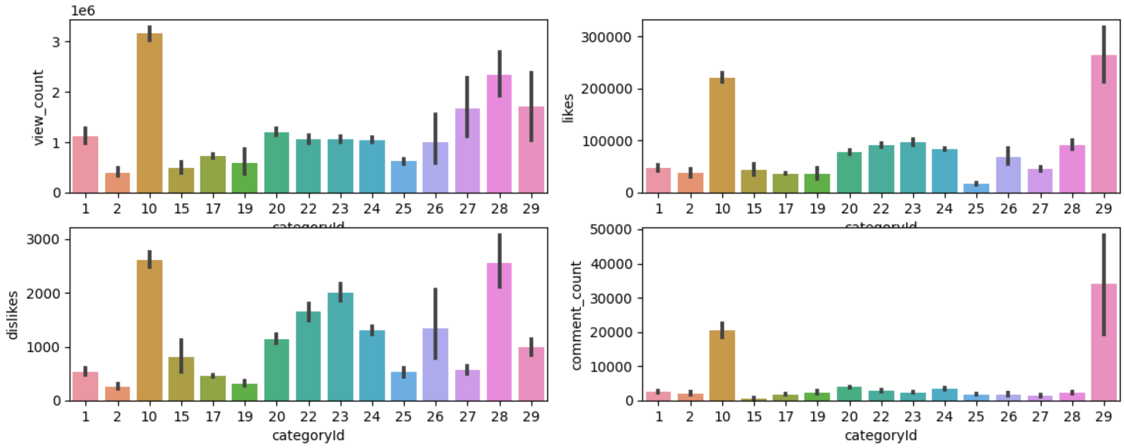

Figure 1.10 : Analysis of Video engagement in Brazil

CANADA:

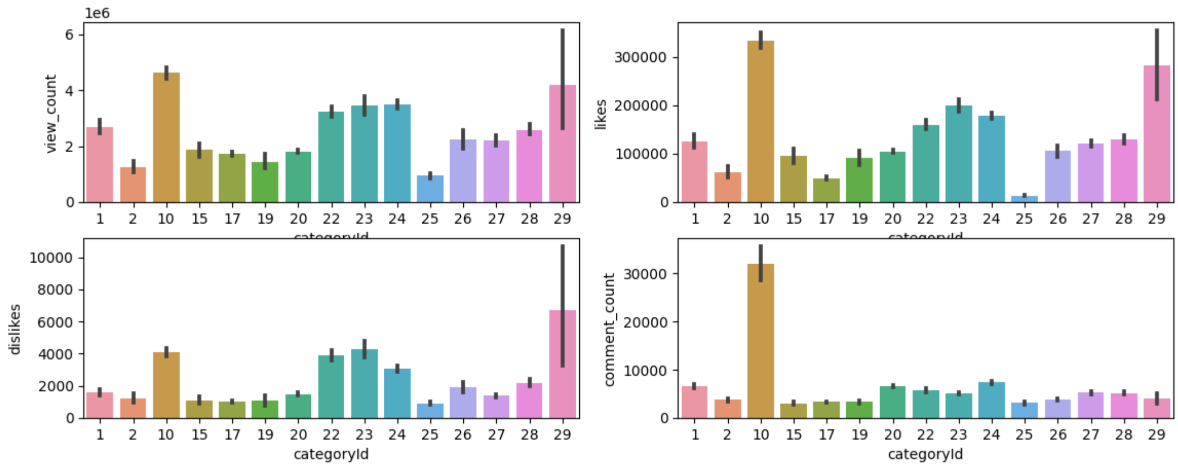

Figure 1.11 : Analysis of Video engagement in Canada

GERMANY:

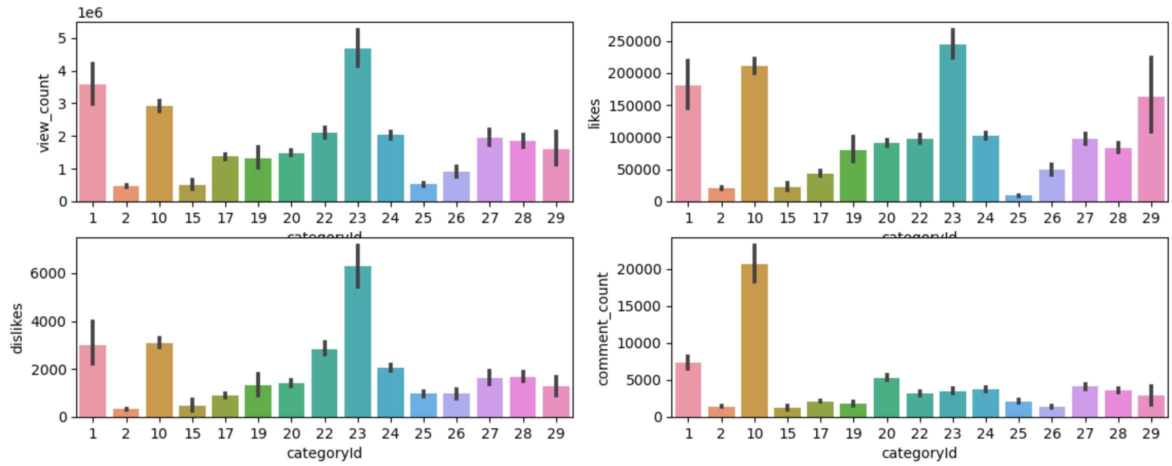

Figure 1.12 : Analysis of Video engagement in Germany

FRANCE:

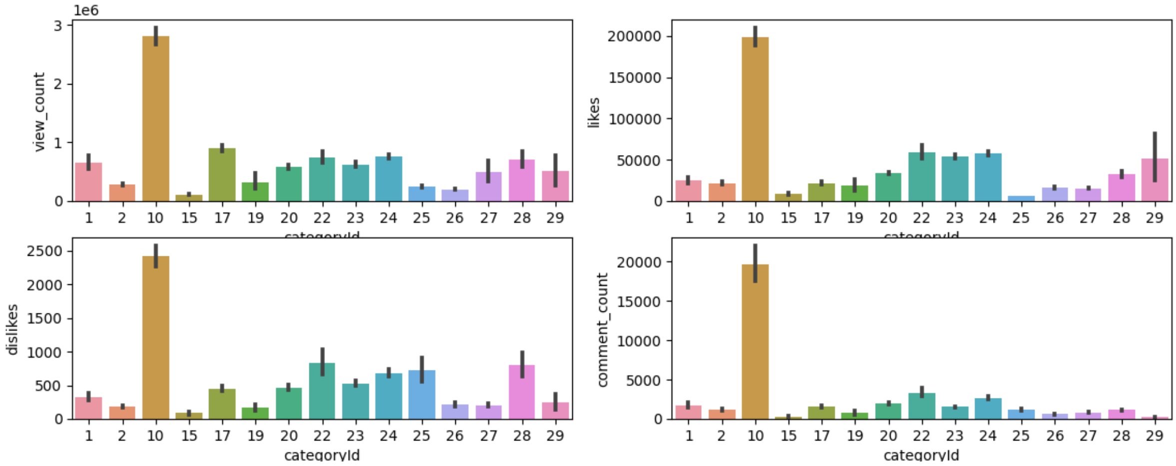

Figure 1.13 : Analysis of Video engagement in France

JAPAN:

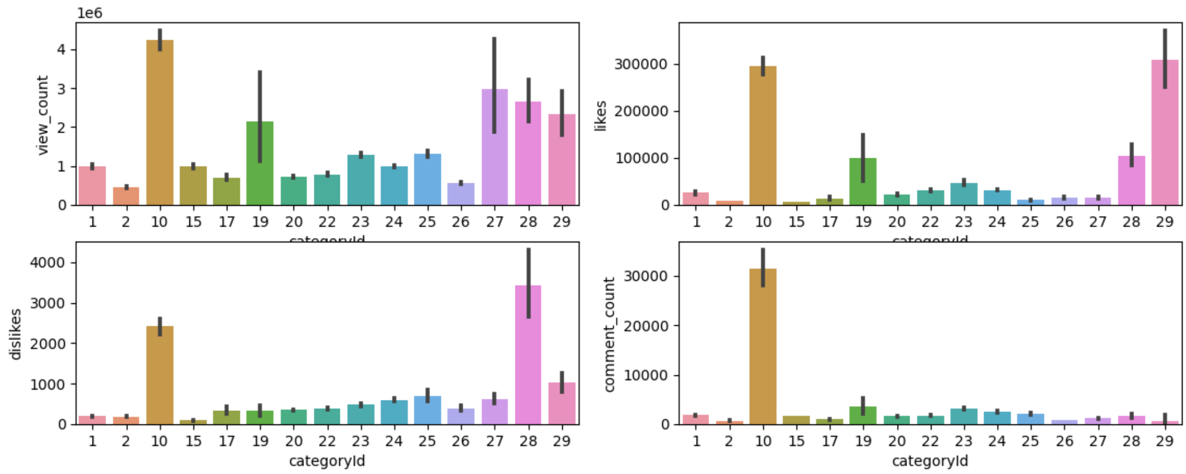

Figure 1.14 : Analysis of Video engagement in Japan

GREAT BRITAIN (GB):

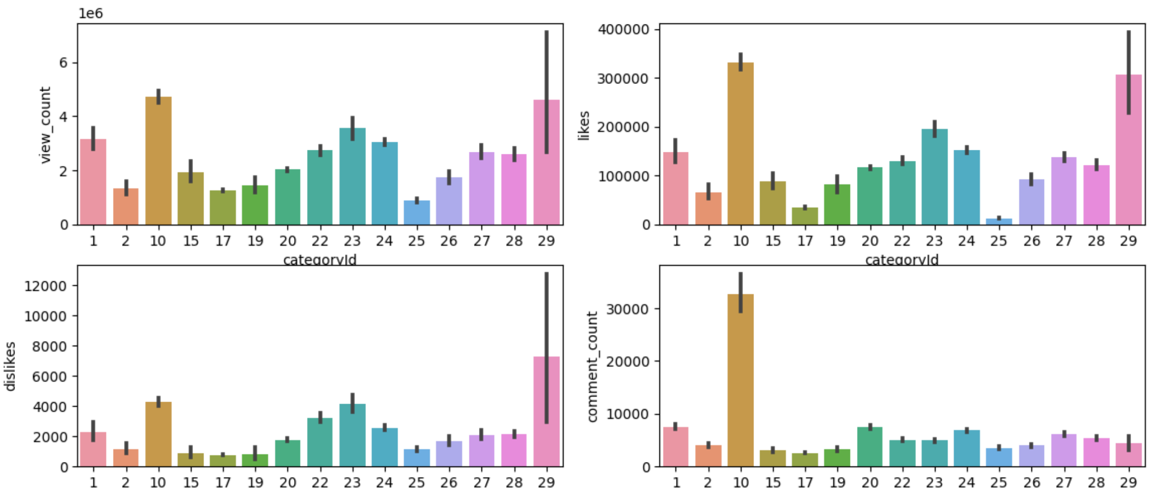

Figure 1.15 : Analysis of Video engagement in Great Britain

Table VI.III : Country wise engagement statistics

Country	Max View Category	Min View Category	Max Like Category	Min Like Category	Max Dislike Category	Min Dislike Category	Max comment	Min comment
India	10	25	10	25	10	23	10	23
Brazil	10	2	29	25	10	2	29	15
Canada	10	25	10	25	29	25	10	15
Germany	23	2	23	25	23	2	10	15
France	10	15	10	25	10	15	10	15
GB	10	25	10	25	29	17	10	17
Japan	10	2	29	15	28	15	10	29
S.korea	10	26	10	2	10	19	10	26

IX. Conclusion:

In majority of the countries, Music was the most watched video category followed by comedy. On the other hand, news and politics were the least engaged video categories. For predicting video views polynomial regression with degree 6 outperformed all the models and knn performed the best for finding video categories with given engagement data.

References:

[i] Asif khan, jian ping , naeem ahmad, shuchi sethi , amin ul haq , sarosh h. patel, and sabit rahim, “Predicting Emerging Trends on Social Media by Modeling It as Temporal Bipartite Networks” ., vol. 8, pp. 39635-39646, 2020

[ii] Amar Krishna, J. Zambreno, Sandeep Krishnan, “Polarity Trend Analysis of Public Sentiment on YouTube” ., pp. 1-42, Iowa State University, 2014

[iii] Chengang zhu , guang cheng , (senior member, ieee), and kun wang(senior member, ieee) , “Big Data Analytics for Program Popularity
Prediction in Broadcast TV Industries” .,vol. 5, pp. 2493-24601, 2017