Analysis of Cross-Country Influence on Trending Youtube Videos

Muskan Goyal(006)
Kritika Rana(014)
Sanya Gupta (028)
Purva Gulati (039)
Muskan Sethi (050)
Kavya Arora (071)

Indira Gandhi Delhi Technical University for Women, Delhi, India
aiproject.team2.2020@gmail.com

Abstract. YouTube facilitates global access to diverse cultural products from all over the world yet consumption of popular videos in culturally different countries appears to be constrained by cultural values. This study unearths the cross-cultural influences on YouTube user engagement by investigating the consumption of trending videos in countries that differ in cultures, language and demographics. We draw on the trending listings maintained by the platform across 10 countries to conduct a systematic and in depth study on the statistics of trending YouTube videos. Analysis was implemented based on user participation, user consumption, categories, tags, titles and region. Results showed that YouTube videos have noticeably different statistics across regions. Popular categories, tags and titles have a strong correlation with region. The results of the analysis also provide insights for content creators and researchers into research trends and issues related to YouTube.

Keywords: YouTube · Cross-cultural · Participation · Consumption · Engagement.

1 Introduction

YouTube is a go-to resource for viewing videos. Launched in 2005, it is the third most visited site in the world after Google and Facebook (Alexa,2016). YouTube content is diverse and global, offering the opportunity to disseminate content to a very broad audience of site visitors. The site thus serves as an attractive platform for both amateur content creators and media companies alike (Xu,Park,Kim,&Park,2016). Politicians, news organizations, education institutes, businesses, music and film artists, and people from all walks of life use YouTube. Recent years have witnessed the rise of video sharing in various forms. Better broadband Internet speeds and growing mobile device use have also fueled higher video consumption.
According to Pew Research, the use of online video-sharing site showed a constant rise, in which about 33% of US adults had posted a video to an online

site(Anderson,2015). As of July 2015, 400 hours of video content was uploaded every minute on YouTube (Statistica,2015). Amongst various social media platforms, YouTube popularity is right behind Facebook; 77% of Internet users are on Facebook, while 63% use YouTube (Anderson,2015).
YouTube allows users to interact with the site in multiple ways, whereby participation on the site takes a deeper meaning. For example, registered users can rate (like/dislike), upload videos, comment on and share them. This phenomenon has given a greater degree of control to social media users in creating and manipulating content besides creating a sense of community.
Crucial to understanding the future of social media is studying the characteristics that make these sites appealing to people. Such sites are increasingly becoming a single platform for social interaction, information, news, and entertainment. A great deal needs to be learned about why and how users participate and consume information on various online sites. The design of socio-technical systems especially for promoting engagement in terms of maximum user participation is both a theoretical and real-world challenge that researchers strive to understand. This study applies a motivational construct to unpack the motives of user participation and consumption on YouTube. Likes, comments, and shares are common features that enable user participation on social media sites. However, there is a diversity of features that are used for varied reasons. One such feature being cross cultural differences.
YouTube provides access to diverse cultural products from all over the world. While theories suggest that web platforms such as YouTube which provide diverse cultural products facilitate global cultural convergence, it is often seen that consumption of content is constrained by cultural differences and that cross-cultural convergence is more advanced in cosmopolitan countries with cultural values that favor individualism and power inequality.

1.1 Research Objectives

This study is aimed at evaluating the cross-cultural influences on YouTube user engagement by investigating the consumption of trending videos in countries that differ in cultures, language and demographics. The study is implemented based on user participation, user consumption, categories, tags, titles and region.Through exploratory analysis of video statistics, we reveal the various relations between attributes.

The objectives to be achieved through this study are as follows:

1. Evaluate the cross cultural influences on YouTube video statistics.
2. Address why some YouTube videos are globally consumed while others are limited to a single country, despite the existence of a technological infrastructure for global cross-cultural communication.
3. Evaluate the effect of user participation, consumption, video category, tags, title and publish time on video popularity.
4. Construct a prediction model for video popularity and test it for different

configurations.
5. Construct a Text Classification model using Bag of Words and TF-IDF(Term Frequency-Inverse Document Frequency).

2 Related Work

Youtube is one of the most visited websites worldwide,and its rise and to one of the most relevant mass communication media is commendable.
A range of studies has used Youtube data to produce insights into social video sharing. A lot of work focuses on the description and analysis of videos and viewer interaction,the prediction of video popularity, and other factors affecting video usage and sharing.
Various studies have explored the parameters affecting the uploads and views attempting to provide statistical analysis and overall characterization of Youtube [1].
Some studies have worked on Social media engagement and incorporated analysis of specific features on Youtube. [2] and focuses on finding motivation of users to participate and consume information on Youtube at a deeper level by viewing engagement in terms of liking , disliking , commenting, sharing, uploading, viewing, and reading comments. [3]focuses on a systematic and in-depth measurement study on the statistics of YouTube videos and the parameters which makes this platform a popular one. [4] is another study which analyses the engagement behaviour in context with video creators to promote video marketing. While there are many studies related to user engagement on YouTube videos consisting of working with attributes like views,likes, tags,categories, duration etc. associated with the videos, the parameter of region also plays an important role in making a video trending. What is lacking is a more global quantitative description of YouTube, focused on global as well as region-wise parameters affecting the popularity of a video.

3 Methodology

3.1 Dataset Description

Source: `https://www.kaggle.com/`
Dataset: `https://www.kaggle.com/datasnaek/youtube-new`
Method adopted for collection of this data:This dat aset was collected using the YouTube API
Table 1 describes the data set through counts of some key entities involved in the data set

Details	Description
Number of instances	37352
Number of attributes	16
Whether labeled or unlabeled	Unlabeled
Type of label information (if present)	N/A
Number of unique videos	24427

Table 1. Details of the dataset.

The data set comprises of 16 data attributes.
Table 2 describes attributes of data.

Data Attributes	Brief Explanation
Video ID(Numeric)	Unique ID to identify the video
Trending date(Date)	Date on which video made it to the trending list
Title(Categoric)	Name of the video
Channel$_t itle(Categoric)$	Name of channel which posted the video
Category$_i d(Numeric)$	Specific to region
Publish$_t ime(Date)$	Time at which video was uploaded
Tags(Categoric)	Words and phrases used to give YouTube context about a video
Views(Numeric)	Number of views on the video
Likes(Numeric)	Number of likes on the video
Dislikes(Numeric)	Number of dislikes on the video
Comment$_c ount(Numeric)$	Number of comments on the video
Thumbnail$_l ink(Categoric)$	Thumbnail link of YouTube video
Comments$_d isabled(Boolean)$	Whether to or not disable comment option
Ratings$_d isable(Boolean)$	Whether to or not disable rating option
Video$_e rror_o r_r emoved(Boolean)$	Either error in playing video or video has been removed from the platform
Description(Categorical)	Brief description of the video

Table 2. Details of Data Attributes.

3.2 Data Pre-processing

3.2.1 Data Files Combining Data was available in the form of 10 .csv files, each file holding the data for one country. These files were combined to form one dataset, with the addition of an addition attribute - 'region'.

3.2.2 Data Transformations Data of all attributes was of the type : string. The following changes were done -
1. All numeric columns (Category ID,Views,Likes,Dislikes,Comment Count) was transformed into type : int.
2. Trending Date column was transformed into datetime format.
3. Publish Time column (consisting of both publish date and publish time) was divided into 2 columns : Publish Date and Publish Time , in the datetime format.

3.3 Proposed Approach

First we focus on exploratory data analysis to understand the data. Focus was on analyzing data sets to summarize their main characteristics using visual methods. EDA was done for seeing what the data can tell us beyond the formal modeling or hypothesis testing task.

Further, we use the concept of Linear Regression to analyze trends across countries.

Algorithm 1 Linear Regression Algorithm

procedure (N)
  for i=1 to n **do** read xi,yi $i \leftarrow i + 1$
  end for
$sumX \leftarrow 0$ $sumX2 \leftarrow 0$ $sumY \leftarrow 0$ $sumXY \leftarrow 0$
  for i=1 to n **do**
sumX=sumX+Xi
sumX2=sumX2+Xi*Xi
sumY=sumY+Yi
sumXY=sumXY+Xi*Yi
$i \leftarrow i + 1$
  end for
For a and b of y = a + bx:
b = (n * sumXY - sumX * sumY)/(n*sumX2 - sumX * sumX)
a = (sumY - b*sumX)/n
Print a,b
end procedure

Lastly, text classification using Naive Bayes was performed. Text classification is the process of assigning tags or categories to text according to its content. Text classification structures text in a fast and cost-efficient way to enhance decision-making and automate processes.

4 Exploratory Data Analysis

The analysis of the data has been done through exploration of the attributes. The aim is to establish patterns, check hypothesis and to spot anomalies with the help of summary statistics and visualisation.
Exploratory Data Analysis is carried out for each attribute

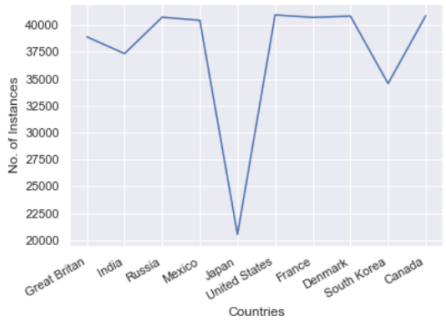

Fig. 1. Number of instances in each region

Figure 1 gives us an idea of number of instances in each region.

4.1 Views, Likes, Dislikes and Comments

In this subsection, the mean values for the attributes are analysed and it's country-wise variation.
Figure 2 depicts the mean values of views, likes, dislikes, comment count and their variation across different regions. From this analysis,we can conclude that: Majority of the countries have Average views value of less than or equal to 1 million views. Majority of the countries have Average Likes value of less than or equal to 50,000 Likes. Majority of the countries have Average Dislikes values less than or equal to 2000 Dislikes. Majority of the countries have Average Comment count value of less than or equal to 4,000 comments.

Fig. 2. Mean values of attributes across different regions

4.2 Time to Trend (trending_time - publish_time)

The time taken for a video to trend depends on the publish time and the trending time. Analysis of the number of days it takes for a video to trend varies country-wise is done. From Figure 3, it can be derived that on an average it takes 7.5 days for a video to trend.It is also noted that videos from Great Britain takes the maximum time to trend followed by United States.

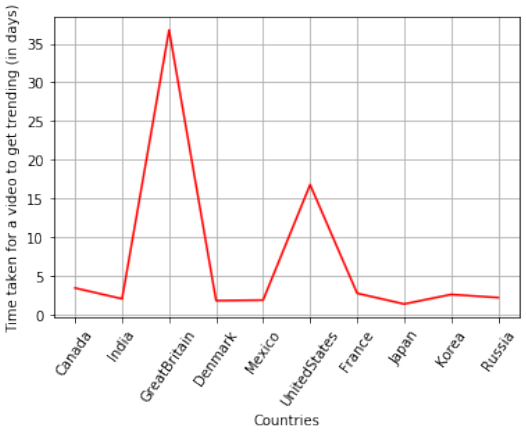

Fig. 3. Average time taken for a video to get trending

Correlation Between Attributes Region-wise

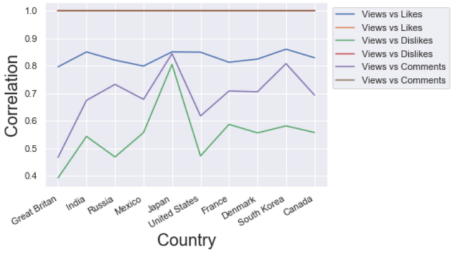

Fig. 4. Correlations

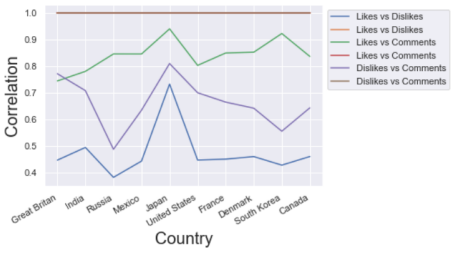

Fig. 5. Correlations

From Figure 4 and 5, the various correlation parameters for each region can be analysed .The values vary for each region. We can observe that there is a strong positive correlation between the number of likes and the number of trending videos: As one of them increases,and vice versa. There exists a strong positive correlation also between number of views and the number of likes. Further, we can conclude that United States has the maximum value in the majority of correlations.
Overall for all the countries, the correlation value between likes and comments, views and likes is the highest.

4.3 comments_disabled, ratings_disabled

In this subsection, analysis of how user participation affects views is done. Comparison of the difference of number of views when the comments are enabled/disabled and ratings are enabled/disabled is done.
Figure 6 shows the comparison of number of views on the trending videos when comments are enabled/disabled. From the graph,one can make out the videos for which comments are enabled get more views.
Figure 7 shows the comparison of number of views on the trending videos when ratings are enabled/disabled. From the graph,one can make out the videos for which ratings are enabled get more views.
Videos with comments enabled and rating enabled yield more views which verifies our hypothesis that a trending video is a direct result of user participation.

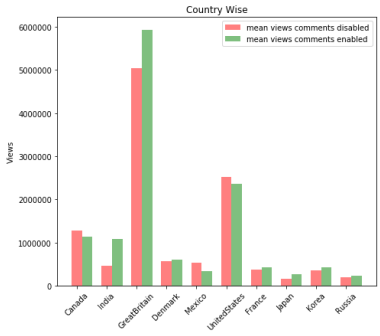

Fig. 6. Comparison of number of views when comments are enabled/disabled

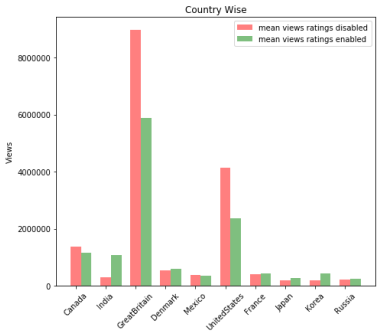

Fig. 7. Comparison of number of views when ratings are enabled/disabled

4.4 Publish Time

Analysis of publish time is done to identify the time slot with the highest number of uploads and to see how publish time affects user participation and user engagement.

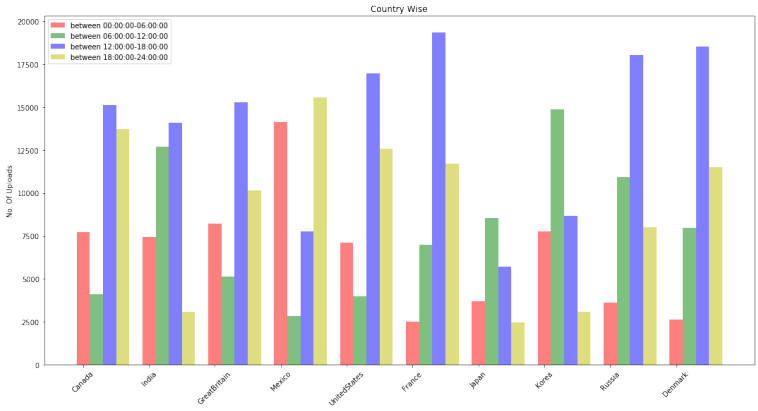

Fig. 8. Number of videos published in different time across different regions

According to the analysis done in Figure 8,the most popular publish time slot is between 12:00 to 18:00

5 Linear Regression

The data was analysed for convenience of regression analysis.
Univariate normality assumption was analysed for each quantitative variable by examining probability distribution function . The PDF shows how that variable is distributed which makes it easy to spot anomalies such as outliers. On the basis of the PDFs, features(variable) to be transformed were identified as likes, dislikes, views and comment_count and the outliers were dealt with.
For the multivariate normality and linearity assumption, the scatter matrix is examined. Examination revealed that no linear relation existed which was dealt with by performing logarithmic transformations.

Multiple Linear Regression model is defined as
$y_i = \beta_0 + \beta_1 x_i 1 + \beta_2 x_i 2 + ... + \beta_p x_i p + \epsilon$
where, for i = n observations:
y_i =dependent variable
x_i = explanatory variables
β_0 = y-intercept (constant term)
β_p =slope coefficients for each explanatory variable
ϵ =the model's error term (also known as the residuals)

Four models are taken :
MODEL 1
Target = views
Inputs = likes,dislikes,comment count, time to trend,publish hour, category ID, tags count , comments disabled, ratings disabled ,video error or removed

MODEL 2
Target = likes
Inputs = views,dislikes,comment count, time to trend,publish hour, category ID, tags count , comments disabled, ratings disabled ,video error or removed

MODEL 3
Target = dislikes
Inputs = views,likes,comment count, time to trend,publish hour, category ID, tags count , comments disabled, ratings disabled ,video error or removed

MODEL 4
Target = Comments
Inputs = views,likes,dislikes, time to trend,publish hour, category ID, tags count , comments disabled, ratings disabled ,video error or removed

5.1 Global Linear Regression

R-squared - measures the proportion of the variation in your dependent variable (Y) explained by your independent variables (X) for a linear regression model.

Regression Coefficients(also called weights) - Higher the absolute value of weight(regression coefficient) of a feature , higher is its influence on the target.

5.1.1 Model 1 R- squared = 0.796

	coef	std err	t	P>|t|	[0.025	0.975]
const	11.5969	0.002	7066.466	0.000	11.594	11.600
x1	-0.0528	0.002	-31.674	0.000	-0.056	-0.050
x2	5.982e-17	4.16e-19	143.872	0.000	5.9e-17	6.06e-17
x3	-1.094e-16	1.61e-18	-67.853	0.000	-1.13e-16	-1.06e-16
x4	0.0070	0.002	4.157	0.000	0.004	0.010
x5	0.1861	0.002	100.875	0.000	0.182	0.190
x6	-0.0809	0.002	-49.034	0.000	-0.084	-0.078
x7	0.0128	0.002	7.657	0.000	0.010	0.016
x8	0.5223	0.004	147.852	0.000	0.515	0.529
x9	1.0099	0.003	362.590	0.000	1.004	1.015
x10	-0.0620	0.003	-17.714	0.000	-0.069	-0.055

$y = 11.59 - 0.05x_1 + 0x_2 - 0x_3 + 0.007x_4 + 0.18x_5 - 0.08x_6 + 0.01x_7 + 0.52x_8 + 1.00x_9 - 0.06x_10$

The regression model reveals that views are highly influenced by dislikes whereas comments_disabled and ratings_disabled have no influence on the number of views a video receives.

5.1.2 Model 2 : R - squared : 0.804

	coef	std err	t	P>|t|	[0.025	0.975]
const	7.6670	0.002	4155.519	0.000	7.663	7.671
x1	-0.0707	0.002	-37.748	0.000	-0.074	-0.067
x2	3.056e-16	6.18e-19	494.552	0.000	3.04e-16	3.07e-16
x3	-7.285e-17	1.33e-18	-54.967	0.000	-7.54e-17	-7.02e-17
x4	-0.0080	0.002	-4.256	0.000	-0.012	-0.004
x5	0.0700	0.002	33.039	0.000	0.066	0.074
x6	0.0467	0.002	25.081	0.000	0.043	0.050
x7	0.0505	0.002	26.961	0.000	0.047	0.054
x8	0.5754	0.004	147.852	0.000	0.568	0.583
x9	0.0560	0.004	14.045	0.000	0.048	0.064
x10	1.1456	0.003	375.501	0.000	1.140	1.152

$y = 7.66 - 0.07x_1 + 0x_2 - 0x_3 - 0.008x_4 + 0.07x_5 + 0.04x_6 + 0.05x_7 + 0.57x_8 + 0.05x_9 + 1.14x_10$

The regression model reveals that likes are highly influenced by the number

of comments whereas comments_disabled and ratings_disabled have no influence on the number of likes a video receives.

5.1.3 Model 3 R - squared = 0.786

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.6535	0.002	2675.024	0.000	4.650	4.657
x1	0.0742	0.002	42.051	0.000	0.071	0.078
x2	-2.166e-17	5.83e-19	-37.163	0.000	-2.28e-17	-2.05e-17
x3	5.792e-17	1.46e-18	39.776	0.000	5.51e-17	6.08e-17
x4	-0.0031	0.002	-1.738	0.082	-0.007	0.000
x5	-0.0365	0.002	-18.250	0.000	-0.040	-0.033
x6	0.0294	0.002	16.746	0.000	0.026	0.033
x7	0.0041	0.002	2.336	0.019	0.001	0.008
x8	1.0979	0.003	362.590	0.000	1.092	1.104
x9	0.0552	0.004	14.045	0.000	0.048	0.063
x10	0.4965	0.004	139.792	0.000	0.490	0.504

$y = 4.65 + 0.07x_1 - 0x_2 + 0x_3 - 0.003x_4 - 0.03x_5 + 0.02x_6 + 0.004x_7 + 1.09x_8 + 0.05x_9 + 0.49x_10$

The regression model reveals that dislikes are highly influenced by the number of views whereas comments_disabled and ratings_disabled have no influence on the number of dislikes a video receives.

5.1.4 Model 4 R- squared = 0.781

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.6582	0.002	3098.045	0.000	5.655	5.662
x1	0.0311	0.002	16.746	0.000	0.027	0.035
x2	3e-16	1.61e-18	186.700	0.000	2.97e-16	3.03e-16
x3	3.597e-17	9.33e-19	38.571	0.000	3.41e-17	3.78e-17
x4	0.0027	0.002	1.435	0.151	-0.001	0.006
x5	0.0051	0.002	2.429	0.015	0.001	0.009
x6	0.0152	0.002	8.248	0.000	0.012	0.019
x7	0.0169	0.002	9.085	0.000	0.013	0.021
x8	-0.0716	0.004	-17.714	0.000	-0.080	-0.064
x9	1.2015	0.003	375.501	0.000	1.195	1.208
x10	0.5278	0.004	139.792	0.000	0.520	0.535

$y = 5.65 + 0.03x_1 + 0x_2 + 0x_3 + 0.002x_4 + 0.005x_5 + 0.01x_6 + 0.01x_7 - 0.07x_8 + 1.2x_9 + 0.52x_10$

The regression model reveals that the number of comments are highly influenced by the number of likes whereas comments_disabled and ratings_disabled have no influence on the number of comments a video receives.

5.2 Country Wise Linear Regression

Country specific Regression is performed to see how influences of features on video popularity vary across countries.

Inputs - comments disabled , ratings disabled , video error or removed , tag counts - these features have Regression Coefficients close to 0.0 , hence they have negligible influence on target values in all 4 models, hence removed from all 4 models.
Train-Test Split - 50:50

5.2.1 Comparison of Regression Coefficients of Inputs This study compares Regression Coefficients of all features, country-wise.

Model 1

From fig 9 and 10 , it is evident that Great Britain has the highest value for Likes Regression Coefficient, hence number of likes has the largest influence on number of views in Great Britain.
France has the highest value for Dislikes Regression Coefficient.
Japan & South Korea have the highest value for Comment Count Regression Coefficient.
Mexico has the highest value for TimeToTrend Regression Coefficient.
India has the highest value for Category ID Regression Coefficient.
Denmark has the highest value for Publish Hour Regression Coefficient.

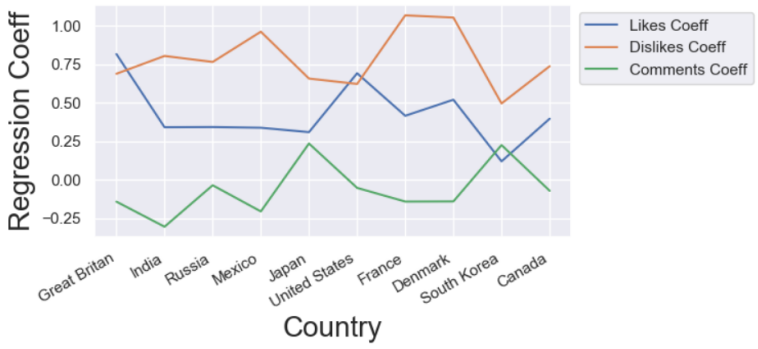

Fig. 9. Regression Coeff. for Likes, Dislikes , Comment Count

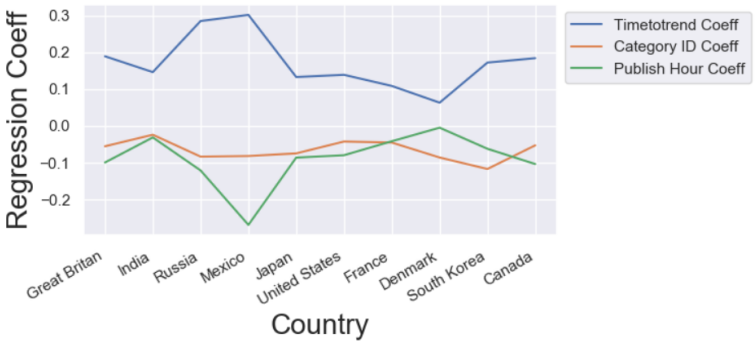

Fig. 10. Regression Coeff. for TimeToTrend , Category ID , Publish Hour

Model 2

From fig 11 and 12, it is evident that Great Britain has the highest value for Views Regression Coefficient, hence number of views has the largest influence on number of likes in Great Britain.
France has the highest value for Comments Regression Coefficient.
Japan & South Korea have the highest value for Comments Regression Coefficient.
Mexico has the highest value for Publish Hour Regression Coefficient.
India has the highest value for Comments Regression Coefficient.
Denmark has the highest value for Comments Regression Coefficient.

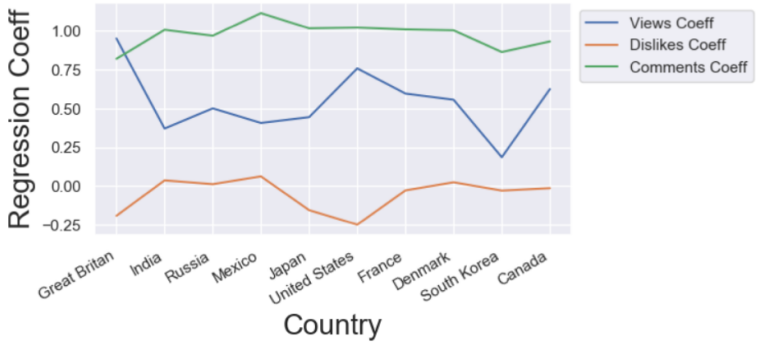

Fig. 11. Regression Coeff. for Likes, Dislikes , Comment Count

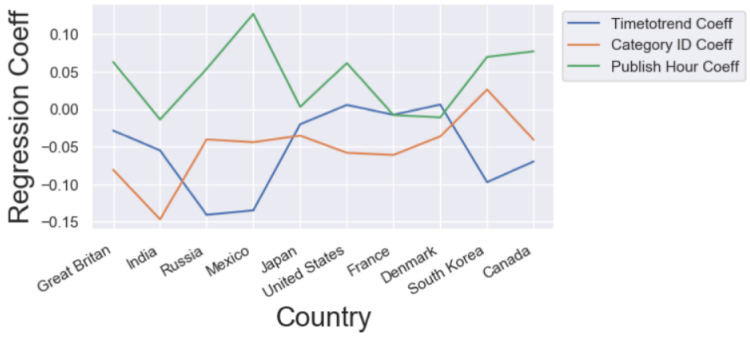

Fig. 12. Regression Coeff. for TimeToTrend , Category ID , Publish Hour

Model 3

From fig 13 and 14 , it is evident that Great Britain has the highest value for Views Regression Coefficient, hence number of views has the largest influence on number of dislikes in Great Britain.
France has the highest value for Views Regression Coefficient.
Japan & South Korea have the highest value for Category ID Regression Coefficient.
Mexico has the highest value for Views Regression Coefficient.
India has the highest value for Views Regression Coefficient.
Denmark has the highest value for Views Regression Coefficient.

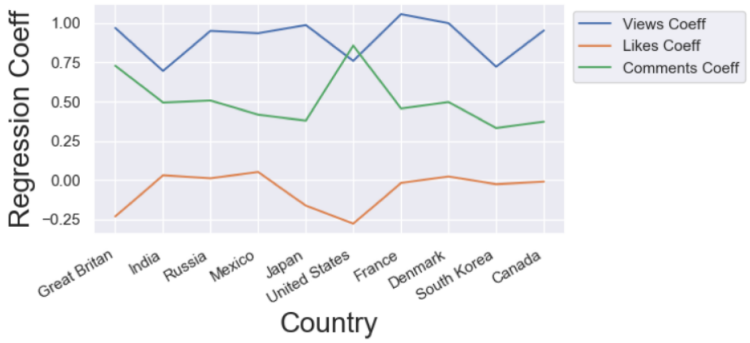

Fig. 13. Regression Coeff. for Likes, Dislikes , Comment Count

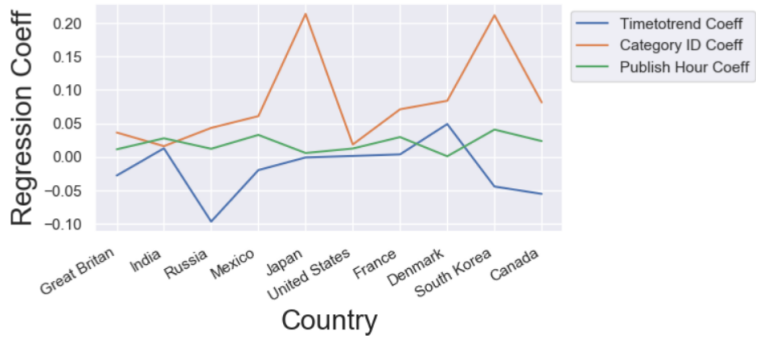

Fig. 14. Regression Coeff. for TimeToTrend , Category ID , Publish Hour

Model 4

From fig 15 and 16, it is evident that Great Britain has the highest value for Likes Regression Coefficient, hence number of likes has the largest influence on number of comments in Great Britain.
France has the highest value for Likes Regression Coefficient.
Japan & South Korea have the highest value for Likes Regression Coefficient.
Mexico has the highest value for Likes Regression Coefficient.
India has the highest value for Likes Regression Coefficient.
Denmark has the highest value for Likes Regression Coefficient.

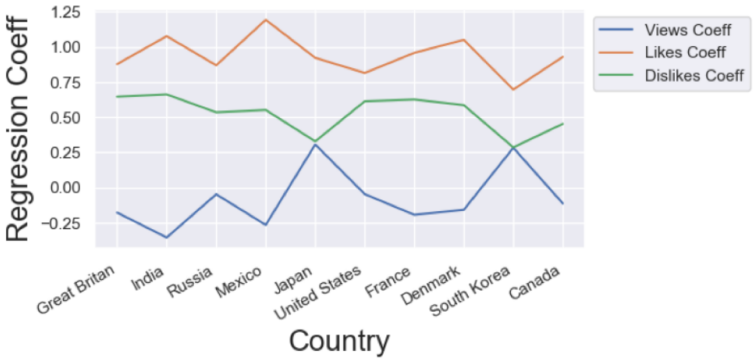

Fig. 15. Regression Coeff. for Likes, Dislikes , Comment Count

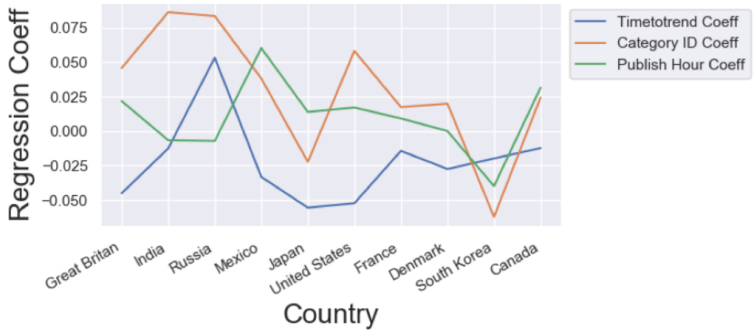

Fig. 16. Regression Coeff. for TimeToTrend , Category ID , Publish Hour

6 Text Classification

Text classification is the process of assigning tags or categories to text according to its content. Text classification structures text in a fast and cost-efficient way to enhance decision-making and automate processes.

In this study we have two text based, descriptive features, tags and titles.

6.1 Tags

Tags are descriptive keywords that are added to videos to help viewers find content close to their search.

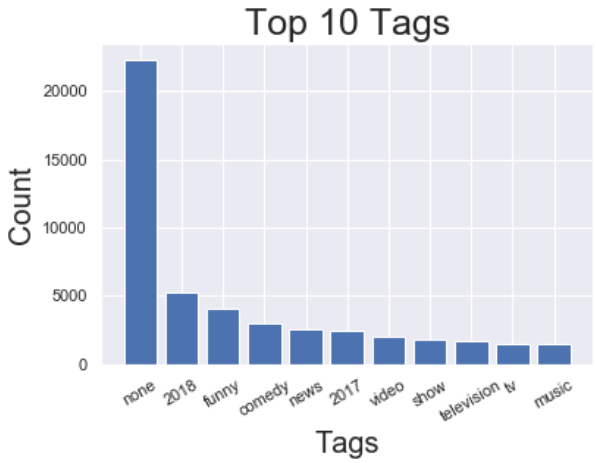

Fig. 17. Top 10 recurring tags

Fig 17 describes the top 10 tags used in trending videos.

In this study, Text Classification is performed for title feature of the data set. Title describes the title description of the YouTube video.

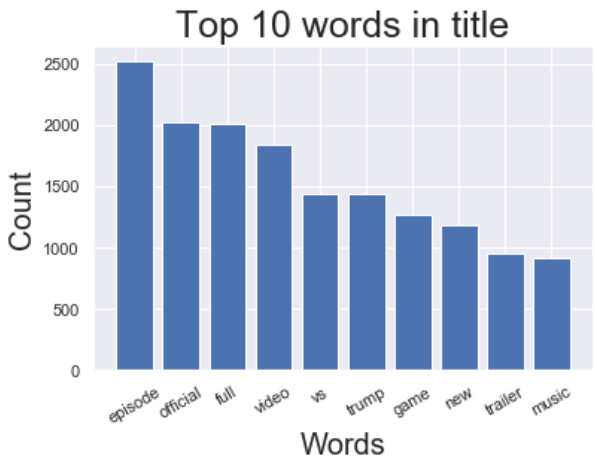

Fig. 18. Top 10 recurring words in title

Fig 18 describes the top 10 recurring words used in trending video titles.

A Multinomial Naive Bayes classifier is used to make classifications.
The first step towards training the classifier is feature extraction: a method is used to transform each text into a numerical representation in the form of a vector.

6.2 Bag of Words

Bag of Words is used for feature extraction.
In Bag of Words, a vector represents the frequency of a word in a predefined dictionary of words.

Five configurations of the training and testing set are evaluated on the basis of MultinomialNB Accuracy.
Fig 19 shows that the Text Classification model using BoW is most accurate for 90:10 configuration.

6.3 TF-IDF

Term Frequency(TF), takes the number of words occurred in each document. The main issue with this Term Frequency is that it will give more weight to

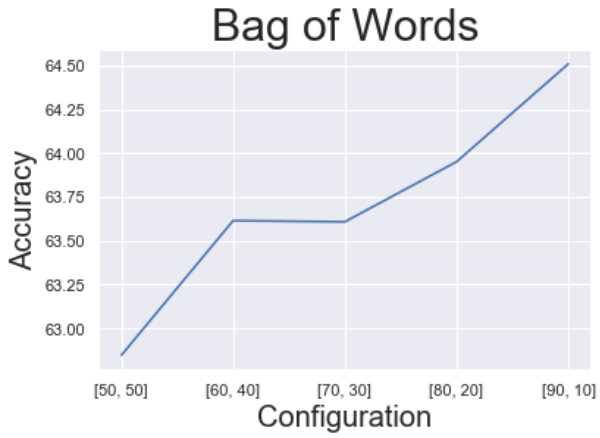

Fig. 19. BoW

longer documents.
IDF(Inverse Document Frequency) measures the amount of information a given word provides across the document. IDF is the logarithmically scaled inverse ratio of the number of documents that contain the word and the total number of documents.
TF-IDF(Term Frequency-Inverse Document Frequency) normalizes the document term matrix. It is the product of TF and IDF.

Five configurations of the training and testing set are evaluated on the basis of MultinomialNB Accuracy.

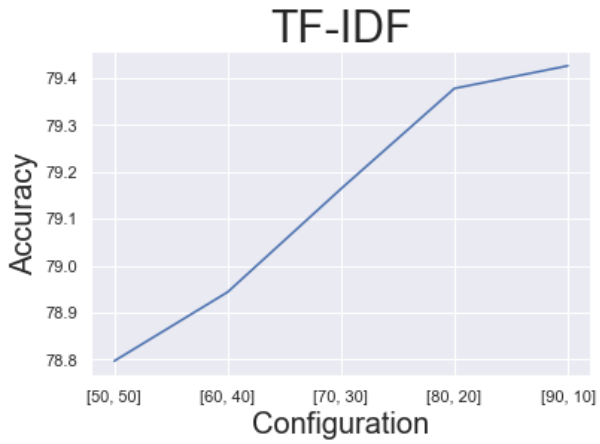

Fig. 20. TF-IDF

Fig 20 shows that the Text Classification model using TF-IDF is most accurate for 90:10 configuration.
Words with high tf-idf in a document are the most occurred words in the given documents and must be absent in the other documents. Hence, these words must be signature words.

6.4 Comparison of Bow and TF-IDF

This study compares Text Classification Model using BoW and TF-IDF for different configurations of training and testing split against MultinomialNB Accuracy.

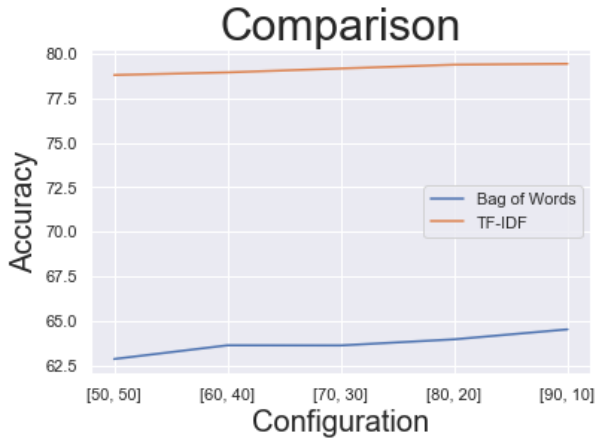

Fig. 21. BoW Vs. TF-IDF

Fig 21 shows that TF-IDF is more accurate than Bow. This is becuase TF-IDF normalises the document matrix.

7 Appendix

7.1 Results of analysis of the following attributes of a trending YouTube video

There are various user parameters associated with the trending videos.The number of views,number of likes, number of dislikes, number of comments and the time taken for a video to trend differs from video to video.
In this subsection, results from analysis of a trending YouTube video is compiled. Visual Analysis of the attributes vs Number of videos associated with the value of attribute is done. From (Figure 22), one can note that the majority of the trending videos has 1 million views or less. Further, the majority of trending videos have 50,000 likes or less(Figure 23) and 20,000 dislikes or less(Figure 24). The comment count in majority of the trending videos is 4,000 or less(Figure 25). By analyzing the time taken to trend after a video is published,it is noted that the majority of the trending videos take one day to get trending(Figure 26).

Visual Analysis of all attributes

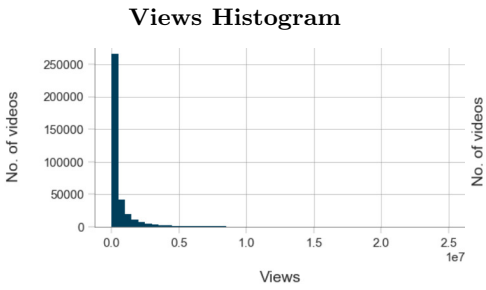

Fig. 22. Number of views associated with videos

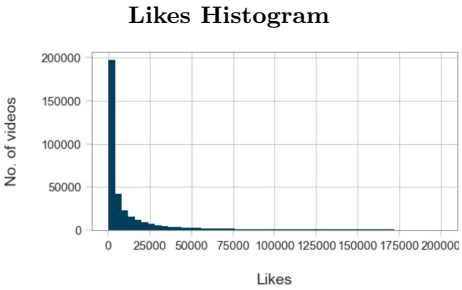

Fig. 23. Number of likes associated with videos

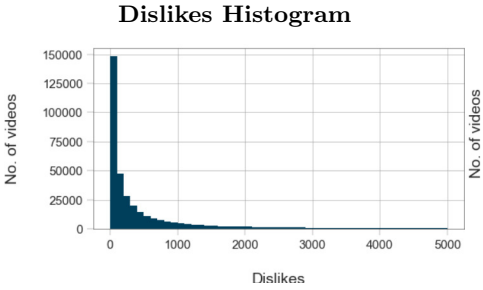

Fig. 24. Number of dislikes associated with videos

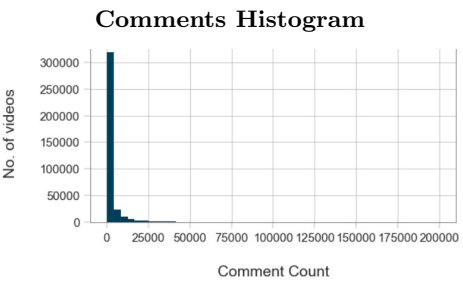

Fig. 25. Comment count associated with videos

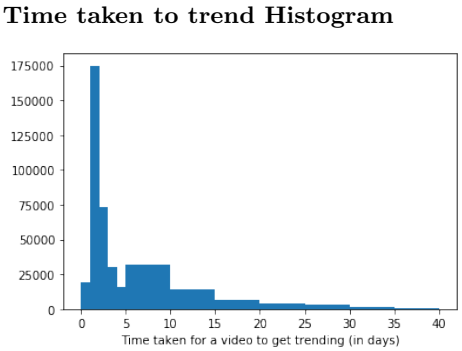

Fig. 26. Time taken by videos to trend

7.2 Country-wise ratio of likes:views, dislikes:views , comments:views

In this subsection, Ratio of Likes:Views, Dislikes:Views, Comments:Views is plotted country-wise
Please note:
1. Likes:views = x means that on average, x of total viewers like the video.
2. Dislikes:views = x means that on average, x of total viewers dislike the video.
3. Comments:views = x means that on average, x of total viewers leave a comment on the video.

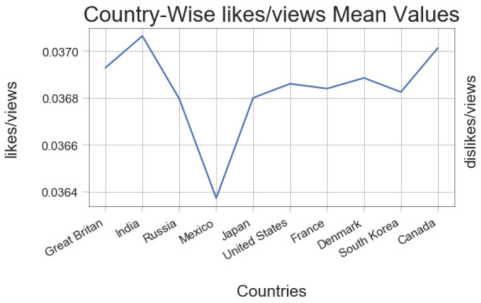

Fig. 27.

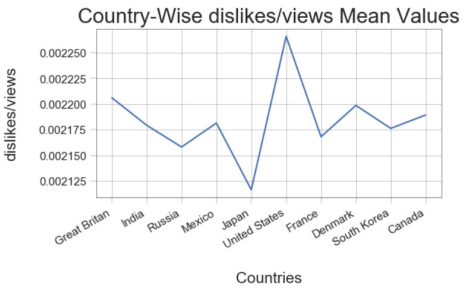

Fig. 28.

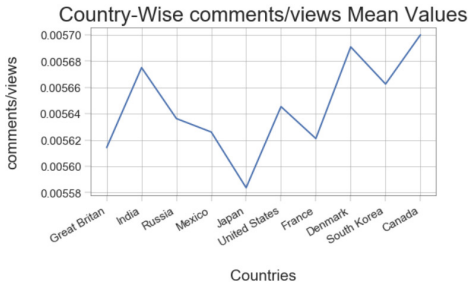

Fig. 29.

According to the figures , India, on an average has the highest ratio of likes/views(more than 0.0370)(Figure 27), US, on an average has the highest ratio of dislikes/views(more than 0.002250)(Figure 28) and Canada, on an average has the highest ratio of comments/views(0.00570)(Figure 29)

7.3 Some results of Global Analysis

Global Analysis of the trending videos are done here. Factors like Upload Time, Likes, Dislikes, Comment Count varies from video to video. In this subsection, the attributes and their values are studied for the trending videos. Visual representations are used to categorise the values of the attributes into groups and interesting results are derived.

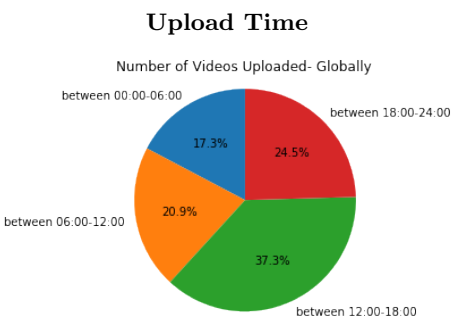

Fig. 30. Upload Time of videos

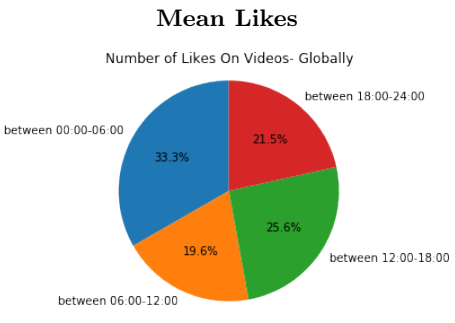

Fig. 31. Likes depending on the publish time

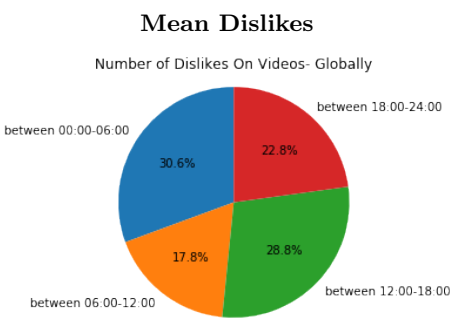

Fig. 32. Dislikes depending on the publish time

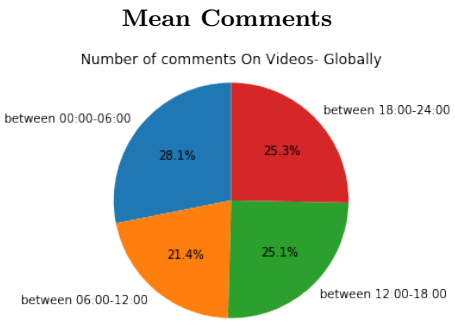

Fig. 33. Comment count depending on the publish time

Analysis of Publish Time of videos-Globally
Figure 30 tells us about the number of videos uploaded in different time slots. From the figure we can note that majority of the videos are uploaded between 12:00-18:00.
Analysis of Likes on videos according to their Publish Time-Globally
Figure 31 tells us about the number of likes on videos which are uploaded in different time slots. From the figure we can note that a video gets maximum likes when was published in the time slot of 00:00-06:00
Analysis of Dislikes on videos according to their Publish Time-Globally
Figure 32 tells us about the number of dislikes on videos which are uploaded in different time slots. From the figure we can note that a video gets maximum dislikes when was published in the time slot of 00:00-06:00
Analysis of Comments on videos according to their Publish Time-Globally
Figure 33 tells us about the number of comments on videos which are uploaded in different time slots. From the figure we can note that a video gets maximum comments when was published in the time slot of 00:00-06:00

7.4 Results on Comments Disabled and Ratings Disabled

In this subsection, focus is on the impact in user engagement when comments and ratings are enabled or disabled.
There are 4 possible cases:
1. Both ratings and comments are disabled
2. Ratings are disabled but comments are enabled
3. Ratings are enabled but comments are disabled.
4. Both ratings and comments are enabled
Analysis of how the number of views and the time taken to trend vary in the given cases is done.

Views Vs. comments disabled and ratings disabled

Figure 34 tells us about the Number of Videos which made to the trending list in the different cases of enabling/disabling of ratings and comments. While Figure 35 gives us the mean views in each case.

After studying the graphs and data set,it can be concluded that since number of instances is much greater for Case 4(When both ratings and comments are enabled), the mean views doesn't give us an accurate picture which may lead to faulty results. In order to deal with that,we consider total views and maximum views. Figure 36 and Figure 37 are used to represent the same.

After studying the data,we can conclude that: Ratio of total number of views on videos with both comments and ratings disabled to total number of views on videos with both ratings and comments enabled = 0.008683448502212093

This implies that, number of views on videos with both comments and ratings disabled is less than number of views on videos with both comments and ratings enabled.

Further,Ratio of max number of views on videos with both comments and ratings disabled to max number of views on videos with both ratings and comments enabled = 0.1468378050584913 This implies that, max number of views on videos with both comments and ratings disabled is less than max number of views on videos with both comments and ratings enabled.

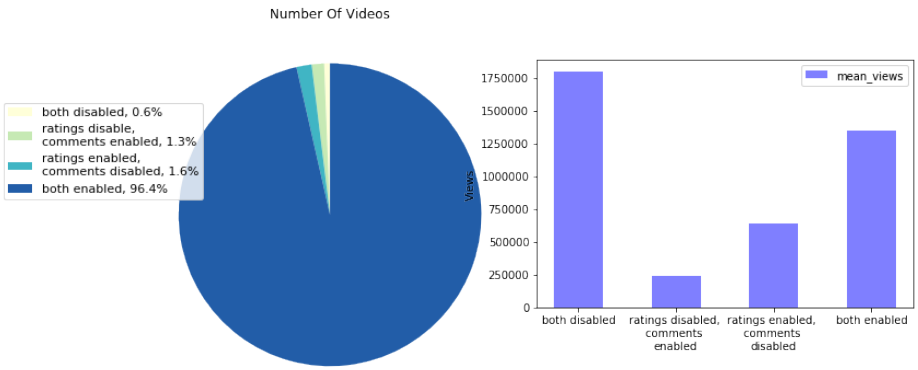

Fig. 34. Number of Videos

Fig. 35. Mean Views for each case

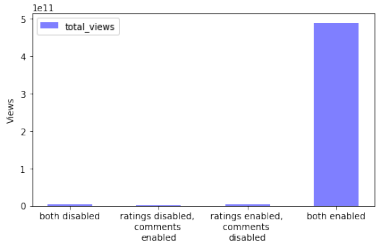

Fig. 36. Total Views

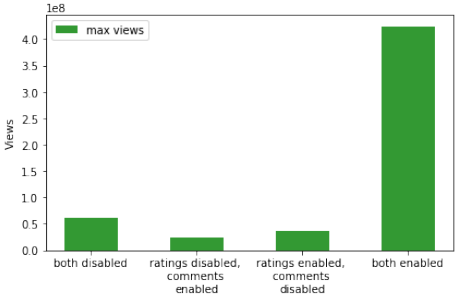

Fig. 37. Max Views

Time to Trend Vs. Comments disabled and Ratings disabled

Time to Trend is the the difference between the date on which the video first got trending and the date on which the video was published. The aim of this analysis is to see the extent to which user participation affects how long it will take for a video to get trending. Figure 38 depicts the same.

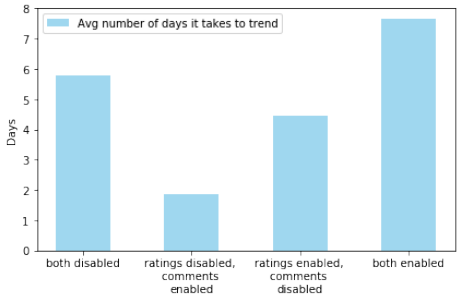

Fig. 38. Average number of days it takes for a video to get trending.

7.5 Analysis of the Title length

Analysis between Title length and Number of views.
A scatter plot between title length and number of views is drawn to see the relationship between these two variables. In this subsection,we find out the relation between the length of the title of the videos that are trending.
From the scatter plot(Figure 39), we can derive that the videos which have 100,000,000 views and more have title length between 33 and 65 characters approximately.
By looking at the scatter plot, it can be said that there is no relationship be-

tween the title length and the number of views.

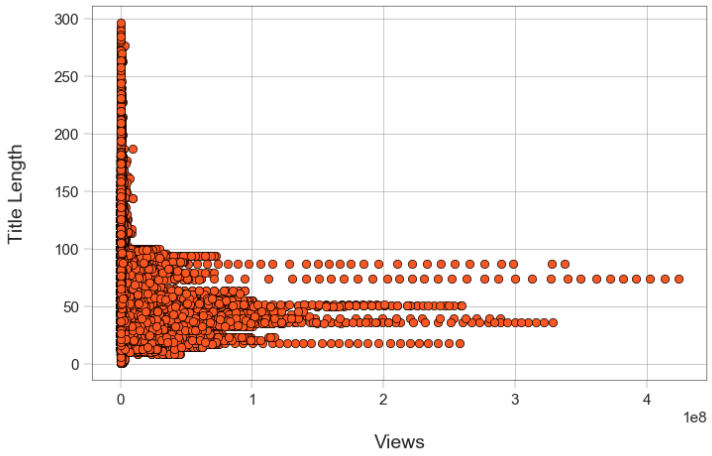

Fig. 39. Number of views vs Title Length

7.6 Categories

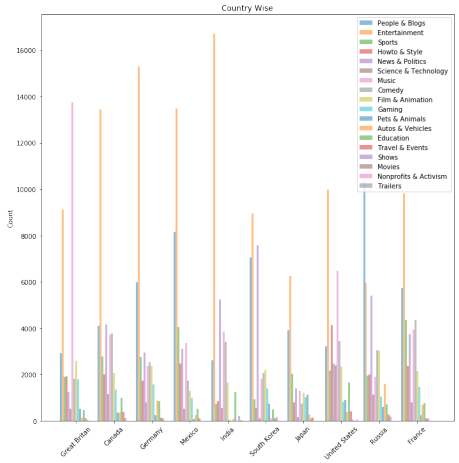

Fig. 40. Division of trending videos in different categories across different regions

We analyse how categories vary country-wise which will help us identify trends which are popular in different regions and also help us see how trends vary across countries.From Figure 40, we can conclude that Entertainment is the most popular category across countries with Great Britain and Russia as exceptions where Music and People Blogs are the most popular categories respectively.

8 Future Work and Conclusion

8.1 Conclusion

YouTube provides access to diverse cultural products from all over the world. While theories suggest that web platforms such as YouTube which provide diverse cultural products facilitate global cultural convergence, it is often seen that consumption of content is constrained by cultural differences and that cross-cultural convergence is more advanced in cosmopolitan countries with cultural values that favor individualism and power inequality.

8.2 Future Work

For future work, Sentiment analysis or opinion mining can be done to analyze opinions, sentiments, evaluations, attitudes, and emotions of users. The online users express their opinions or sentiments on the videos that they watch on such sites.
While this paper presents a brief summary of techniques to analyze opinions posted by users in the form of likes and dislikes, sentiments can also be analysed through comments posted by viewers on videos. The same can be done by extracting comments through the YouTube API.
Future work can also encompass identification of social lexicons through the comments which can help to increase the performance to predict rating of comments.

References

[1]YouTubechannels,uploadsandviews:Astatisticalanalysisofthepast10years-MathiasB\
unhbox\voidb@x\bgroup\let\unhbox\voidb@x\setbox\@tempboxa\hbox{a\global\
mathchardef\accent@spacefactor\spacefactor}\accent127a\egroup\spacefactor\
accent@spacefactorrtl,2018

[2]CialisOnline,Viagra100mgSildenafil~ProfessorkhanOnlineMDStore

[3]https://ieeexplore.ieee.org/abstract/document/4539688/references#
references

[4]DoItfortheViewers!:AudienceEngagementBehaviorsofYoungYouTubers