INDIRA GANDHI
DELHI TECHNICAL UNIVERSITY
FOR WOMEN

Cross Cultural Influences on Trending YouTube Videos

Submitted By

Muskan Goyal - 00601032017
Kritika Rana - 01401032017
Sanya Gupta - 02801032017
Purva Gulati - 03901032017
Muskan Sethi - 05001032017
Kavya Arora - 07101032017

Under the supervision of
Mr. Rishabh Kaushal
Assistant Professor
Department of Information Technology
Indira Gandhi Delhi Technical University for Women

STUDENT UNDERTAKING

Dated: 26.05.2020

This is to undertake that the work titled **Cross Cultural Influences on Trending YouTube Videos** in this Minor Project Report was completed by me as part of 6th Semester in B.Tech. (Information Technology) during January – May,2020.

The report has been written by me in my own words and not copied from elsewhere. This report was submitted to plagiarism detection software on ___ (date) and percentage similarity found was ___, similarity report attached as Appendix.

Anything that appears in this report which is not my original has been duly and appropriately referred / cited / acknowledged. Any academic misconduct and dishonesty found now or in future in regard to above or any other matter pertaining to this report shall be solely and entirely my responsibility. In such a situation, I understand that a strict disciplinary action can be undertaken against me by the concerned authorities of the University now or in future and I shall abide by it.

Student Signature
Muskan Goyal
Kritika Rana
Sanya Gupta
Purva Gulati
Muskan Sethi
Kavya Arora

26.05.2020, New Delhi

DEPARTMENT OF INFORMATION
TECHNOLOGY
INDIRA GANDHI DELHI TECHNICAL
UNIVERSITY FOR WOMEN
KASHMERE GATE, DELHI - 110006

Dated: 26.05.2020

CERTIFICATE

This is to certify that the work titled **Cross Cultural Influences on Trending YouTube Videos** submitted by Muskan Goyal, Kritika Rana, Sanya Gupta, Purva Gulati, Muskan Sethi, Kavya Arora in this project report as part of 6th Semester in B.Tech. (Information Technology) during January – May,2020 was done under my guidance and supervision.

This work is their original work to the best of my knowledge and has not been submitted anywhere else for the award of any credits / degree whatsoever. The work is satisfactory for the award of Minor Project credits.

Mr. Rishabh Kaushal
Assistant Professor
Department of Information Technology
Indira Gandhi Delhi Technical University for Women

ACKNOWLEDGEMENT

We wish to express our sincere gratitude and deepest appreciation towards our teacher, Mr.Rishabh Kaushal, for giving us the opportunity to work on this project titled 'Cross Cultural Influences on Trending YouTube Videos' and his valuable guidance and support throughout the Project.
His continuous instructions, remarks and motivation played a significant role in the completion of this project work.
We also wish to thank our classmates for their support and encouragement. Lastly, we would further like to extend our appreciation to Indira Gandhi Delhi Technical University for Women , for providing our team with an excellent platform to turn our ideas into reality.

Muskan Goyal
Kritika Rana
Sanya Gupta
Purva Gulati
Muskan Sethi
Kavya Arora

Cross-Cultural Influences 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

Abstract. YouTube facilitates worldwide availability of diverse content from all parts of the world. However, it is possible that consumption of videos may vary across different cultures. In this work, we focus on studying the cross-cultural influences on YouTube user engagement by investigating the effects of cultures, language and demographics on the consumption of trending videos in different cultures followed across many countries. 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 significant differences in statistics across cultures. Popular categories, tags and titles have a strong correlation with culture. The outcome of the analysis further provides an understanding for content creators and researchers regarding the trends and problems related to the platform.

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

Table of Contents

1 Introduction

YouTube is a primary platform for publishing and watching videos. Ever since it started in 2005, it has become the third most visited website, after Google and Facebook [1]. YouTube offers a vast variety of content of all genres, to a broad audience from all around the world. The platform hence is a go-to place for content creators and companies to share and advertise their content [6]. YouTube is used by people from all kinds of professions, backgrounds, age groups and nationalities for uploading and viewing content. Lately, there has been a rapid acceleration of sharing content in various forms, for various purposes, due to availability of improved broadband Internet connections and growing use of mobiles.

YouTube permits users to participate on the platform in numerous ways, in the form of ratings(like/dislike), uploading videos, commenting on and sharing them. This ability of participation which enables the users to have control has created a sense of community on the platform.

It is important to examine the features of video sharing platforms which make them attractive to the users, in order to comprehend the future of such platforms. Video sharing platforms are rising as an exclusive place for social interaction, knowledge, amusement and much more. It is essential to comprehend why and how users engage in content on these platforms. Researchers seek to understand the design of such social media platforms, especially for maximizing engagement of users, which is a real-word and socio-technical challenge [4]. This study aims to explore the motives and reasons of user engagement(both content consumption and participation) on YouTube, on the basis of the global cultural differences. Number of likes, dislikes, comment count and shares are the conventional attributes that allow user participation on video sharing platforms. However, there are a numerous other factors that affect user participation, one such attribute being cross cultural influences.

YouTube facilitates worldwide availability of richly diverse content from all parts of the world. While it is believed 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.

1.1 Research Objectives

Our work focuses on evaluating the cross-cultural influences on YouTube user engagement by investigating the consumption of trending videos in 10 countries that differ from each other in cultures, language and demographics. The study is performed based on user participation, user consumption, categories, tags, titles and region.Through exploratory analysis of video statistics, we find the various latent relations between attributes.

The objectives achieved through this study are:

1. Evaluate the cross cultural influences on YouTube video statistics.

2. Understand why some YouTube videos have a global audience while other videos are specific to a single country,in spite of the having a technological infrastructure that offers 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 that stands for Term Frequency-Inverse Document Frequency.

2 Related Work

Youtube is one of the most popular websites in the world. A range of studies has used Youtube statistics to garner insights into video sharing. Many studies focus on the analysis of videos and interactivity with the viewers, the conjecture of video becoming popular, and various more factors affecting video usage and sharing.
YouTube channels, uploads and views: A statistical analysis of the past 10 years by Mathias BärtlVarious [2] has explored the parameters affecting the uploads and views attempting to provide statistical analysis and overall characterization of Youtube. Social media engagement: What motivates user participation and consumption on YouTube? by M. Laeeq Khan [4] has worked on Social media engagement and incorporated analysis of features of Youtube that are specific to it and focuses on finding motivation of users to indulge and consume data on Youtube by looking at participation in terms of comments,ratings,likes,dislikes,uploads and views.
Statistics and Social Network of YouTube Videos, Chang et al [3] focused on a systematic measurement study on the statistics of YouTube videos and the parameters which makes this platform a popular one. "Do it for the viewers!", McRoberts et al [5] is another study which analysed the engagement behaviour in context with video creators to promote video marketing.
All these studies were related to user engagement on YouTube videos and worked with attributes like views,likes, tags,categories, duration etc. associated with the videos and lacked the parameter of region which plays an important role in making a video trending. We could not find studies which were more universally quantitative and focused on global as well as region-wise parameters affecting the popularity of a video on YouTube.

3 Dataset Description

There are various YouTube statistics datasets available for research. We choose the dataset published by Mitchell J on Kaggle[1].

[1] Source: https://www.kaggle.com/
Dataset: https://www.kaggle.com/datasnaek/youtube-new

Method adopted for collection of this data: This dataset was collected using the YouTube API

Table 1 describes the dataset through counts of some key entities involved in the data set

Table 1. Details of the dataset.

Details	**Description**
Number of instances	37,352
Number of attributes	16
Whether labeled or unlabeled	Unlabeled
Type of label information (if present)	N/A
Number of unique videos	24,427

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

Table 2. Details of Data Attributes.

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_title(Categoric)	Name of channel which posted the video
Category_id(Numeric)	Specific to region
Publish_time(Date)	Time at which video was uploaded
Tags(Categoric)	Words and phrases used to give 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_count(Numeric)	Number of comments on the video
Thumbnail_link(Categoric)	Thumbnail link of YouTube video
Comments_disabled(Boolean)	Whether to or not disable comment option
Ratings_disable(Boolean)	Whether to or not disable rating option
Video_error_or_removed(Boolean)	Either error in playing video or video has been removed from the platform
Description(Categorical)	Brief description of the video

3.1 Data Pre-processing

Data was available in the form of 10 .csv files, each file holding the data for one country. These files are combined to form one dataset, with the addition of an additional attribute - ’region’.

Data of all attributes was of the type : string. Hence, all numeric columns (Category ID,Views,Likes,Dislikes,Comment Count) are transformed into type : int.
Trending Date column is transformed into datetime format and the Publish Time column (consisting of both publish date and publish time) is divided into 2 columns : Publish Date and Publish Time , in the datetime format.
Using publish_date and trending_date, we derive another attribute - 'time_to_trend'

$$time\_to\_trend = trending\_date - publish\_date \tag{1}$$

The trending time has been further divided into four time slots : 00:00-06:00, 06:00-12:00, 12:00-18:00, 18:00-24:00

3.2 Data Analysis

In this section, we do basic analysis of the data in order to proceed with a deeper analysis of the data.

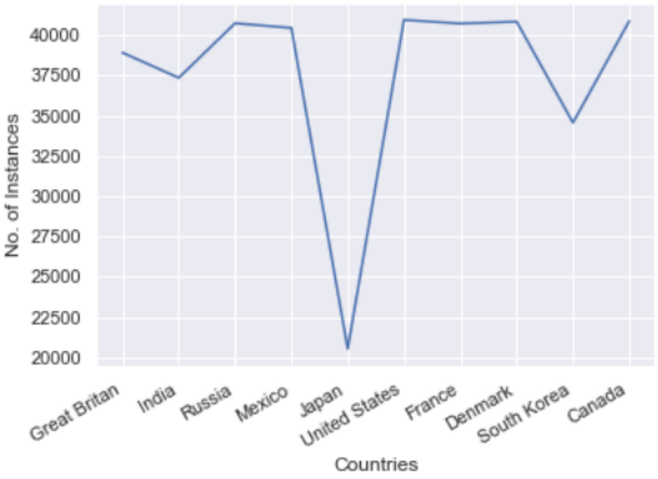

Fig. 1. Number of instances of trending videos in each region. The number of instances is plotted on the y-axis and the countries on the x-axis. We can infer from this graph that the number of instances of Japan is the least

Figure 1 tell us the number of instances that made it to the trending video list on YouTube. From every region we have more than 20,000 videos that made it to the trending list. Japan has the least number of videos that made it to the trending list compared to the other regions.

3.2.1 Views, Likes, Dislikes and Comments

In this subsection, the mean values for the attributes which includes views, likes, dislikes and comments are analysed in the context of the countries.
Figure 2 depicts the mean values of views, likes, dislikes, comment count and their variation across different regions. From this analysis, we conclude that:
(1) Majority of the countries have average views less than or equal to one million views. (2) Majority of the countries have average likes less than or equal to 50,000 likes. (3) Majority of the countries have average dislikes less than or equal to 2,000 dislikes. Majority of the countries have average comment count less than or equal to 4,000 comments.

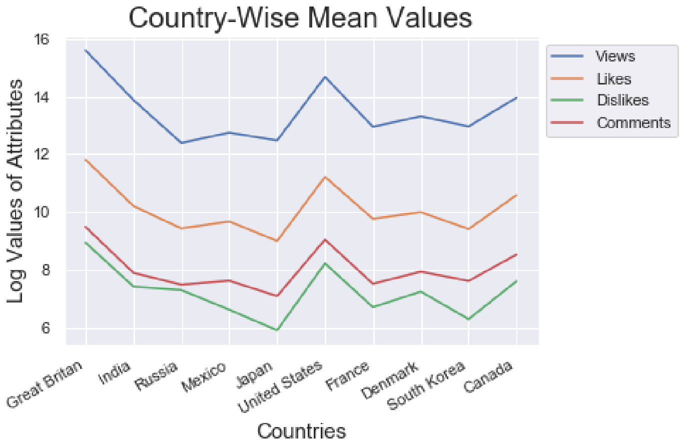

Fig. 2. Mean values of attributes-views, likes, dislikes, comments across different regions. The countries are plotted on the x-axis and the log values of the mean of the attributes on the y-axis.

4 Proposed Methodology

We use machine learning approaches to analyse trends across cultures and for measuring the influences of culture on different YouTube video statistics.
The attribute 'region' is used for describing and contextualising culture.

We divide the problem into two subproblems:
SUBPROBLEM 1. ***Prediction of Numeric Attributes***
Using a linear regression model, we analyse the influences of culture on views, likes, dislikes and comments. We construct four prediction models for our purpose.
SUBPROBLEM 2. ***Classification of Categoric Attributes***
Using a Naive Bayes classifier, we classify video titles based on region.

We solve the two mentioned subproblems in two phases: *prediction of numeric video attributes* and *classification of titles based on region.*

4.1 Prediction of numeric video attributes

In this study, we use a linear regression model to analyse the influences of region on views, likes, dislikes and comments.
Linear Regression models the relationship between 2 or more variables. We have a target variable whose value is predicted by linearly combining the explanatory variables.

First, we analyse the univariate normality assumption for each numeric 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 are identified as likes, dislikes, views and comments and the outliers are dealt with by performing transformations.

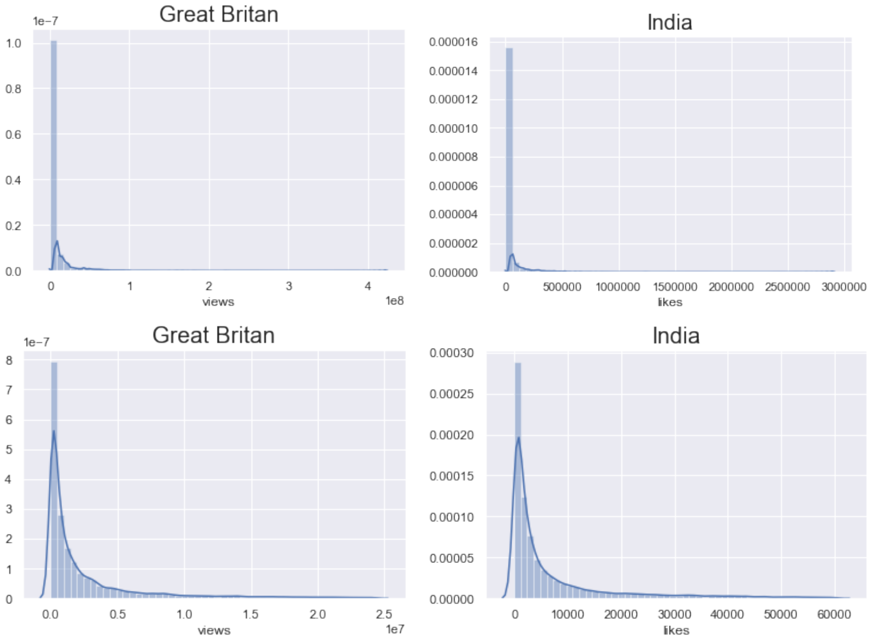

Fig. 3. defines an instance of the probability distribution functions before and after dealing with outliers respectively. From the PDFs we can see that the outliers fall towards the right and hence, we remove 5% of the highest values of the attribute.

For the multivariate normality and linearity assumption, the scatter matrix is examined. Examination reveals that no linear relation exists which is dealt with by performing logarithmic transformations.

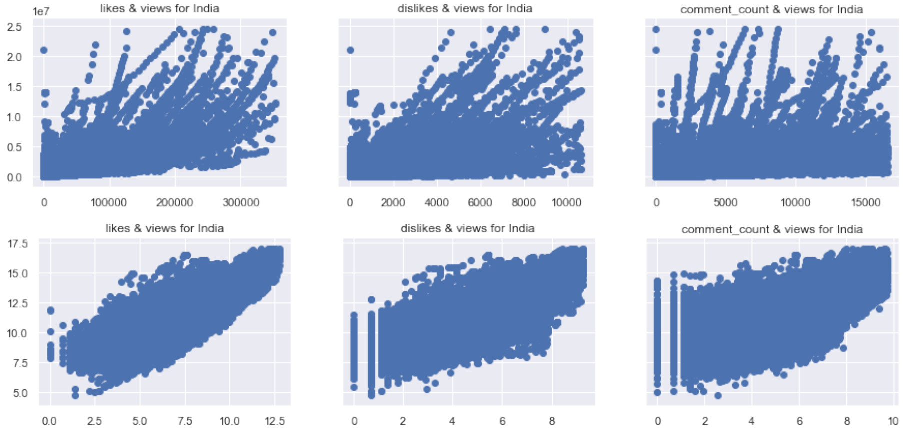

Fig. 4. describes an instance of the scatter plots before and after dealing with outliers respectively.

We define our multiple linear regression model as :
$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + ... + \beta_p x_{i,p}$
where, for i = 4:
Y_i = the dependent variable; $Y_i \in \{viewx, likes, dislikes, comments\}$
$x_{i,j}$ = the explanatory variables;
$x_{i,j} \in \{views, likes, dislikes, comments, time\_to\_trend, publish\_hour,$
$category\_id, tag\_count, comments\_disabled, ratings\_disabled, video\_error\}$
β_0 = y-intercept present(constant term)
β_p = the regression coefficients for each explanatory variable

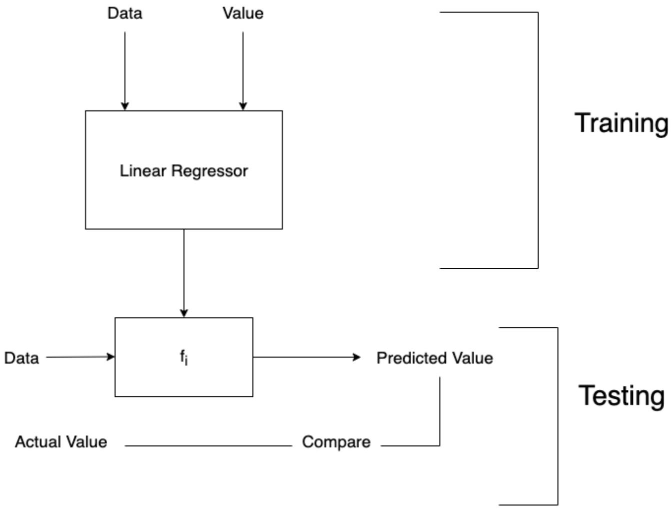

Fig. 5. After pre-processing our data, we split it into training and testing sets. The training set is used for model building. The testing set is used for model evaluation. We test five configurations of the training-testing split, that is, {(50:50), (60:40), (70:30), (80:20), (90:10)}. R-squared is used as an evaluation metric as it measures the proportion of variation present in the dependent variable, Y, which is impacted by the independent variables, Xi, related to the linear regression model.

We consider four linear regressors with different target and explanatory variables.

4.1.1 Prediction of YouTube video views

Target, Y = views
Inputs, X_i = likes, dislikes, comment count, time to trend, publish hour, category ID, tags count, whether are comments disabled or not, whether ratings are disabled or not, video error present or removed

4.1.2 Prediction of YouTube video likes

Target, Y = likes
Inputs, X_i = views, dislikes, comment count, time to trend, publish hour, category ID, tags count, whether are comments disabled or not, whether ratings are disabled or, video error present or removed

4.1.3 Prediction of YouTube video dislikes

Target, Y = dislikes
Inputs, X_i = views, likes, comment count, time to trend, publish hour, cate-

gory ID, tags count, whether are comments disabled or not, whether ratings are disabled or, video error present or removed

4.1.4 Prediction of YouTube video comments

Target, Y = comments
Inputs, X_i = views, likes, dislikes, time to trend, publish hour, category ID, tags count, whether are comments disabled or not, whether ratings are disabled or, video error present or removed

4.2 Classification of YouTube video titles based on region

In this study, we perform text classification using Naive Bayes to classify titles of YouTube videos based on region.
Text classification involves assigning tags or categories to text on the basis of its content. It structures text and enhances decision-making and automate processes.

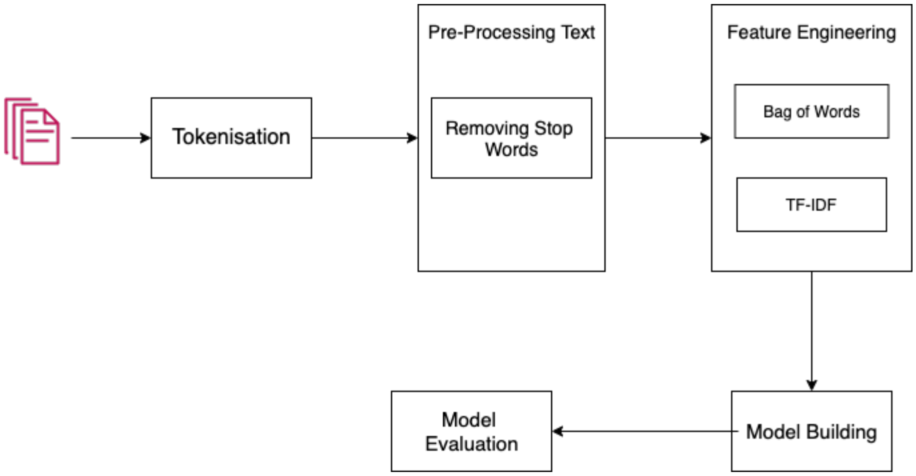

Fig. 6. We break down the video titles into smaller pieces, that is, tokens. This is called tokenisation. The tokens are passed for pre-processing during which all noise in the text, that is, stop words are removed. To begin classifying our data using Naive Bayes, we first perform feature extraction, that is, converting the title into a numeric form. For this purpose we use Bag of Words and TF-IDF.

In BOW model, frequency of a word is represented by a vector.
TF(Term Frequency) takes the count of words occurred in each document and IDF(Inverse Document Frequency) measures the amount of information provided by a across the document. TF-IDF normalises the term matrix of the document. We construct a Naive Bayes Classifier to make classifications.

For model evaluation, we use five configurations of training and testing set, that is, {(50:50), (60:40), (70:30), (80:20), (90:10)}, which are evaluated on the basis of MultinomialNB Accuracy.

5 Evaluation and Results

5.1 Exploratory Data Analysis

We analyse the data through exploration of the various attributes associated with the YouTube videos. The aim is to establish patterns, relationships and to spot anomalies with the help of summary statistics and visualisation.
We carry out Exploratory Data Analysis for each attribute and note the important observations.

5.1.1 Time taken for a video to trend

The time taken for a video to trend depends on the publish time and the trending time. It is the difference between the time when the video became trending and the time the video was published. We analyse the total number of days it will take for a video to trend and it's variation with every country. From Figure 7, we derive that on an average it takes 7.5 days for a video to trend. We note that videos from Great Britain takes the maximum time to trend followed by United States.

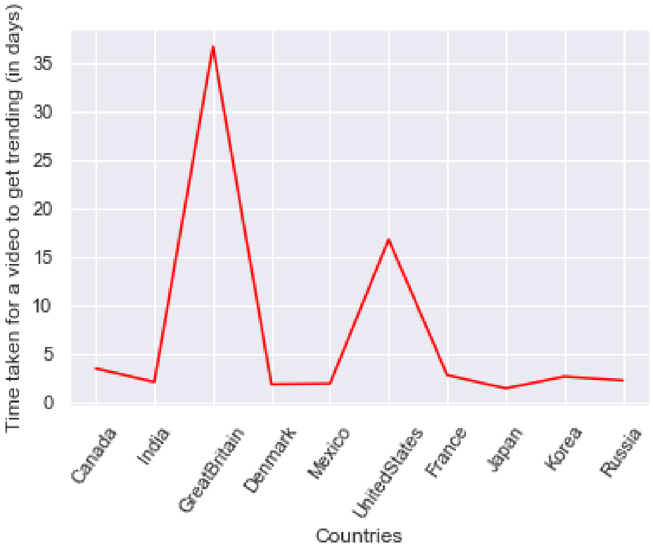

Fig. 7. Average time taken for a video to get trending (in days) is plotted on the y-axis and the countries are plotted on the x-axis. We can infer from this analysis that on an average 7.5 days is taken for a video to trend

5.1.2 Region-wise Correlation Between Attributes

From Figure 8, the various correlation parameters for each region can be analysed. The values vary for each region. We can observe that there exists a really strong positive correlation between the parameters of count of likes and the count of trending videos: As one parameter increases, the other will increase too and vice versa. Similarly, there exists a fairly strong positive correlation between the count of views on the videos and the count of likes accounted. 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.

It is also worth noting that the correlation between views and likes remains almost steady across all regions. This implies that user participation prompts user consumption.

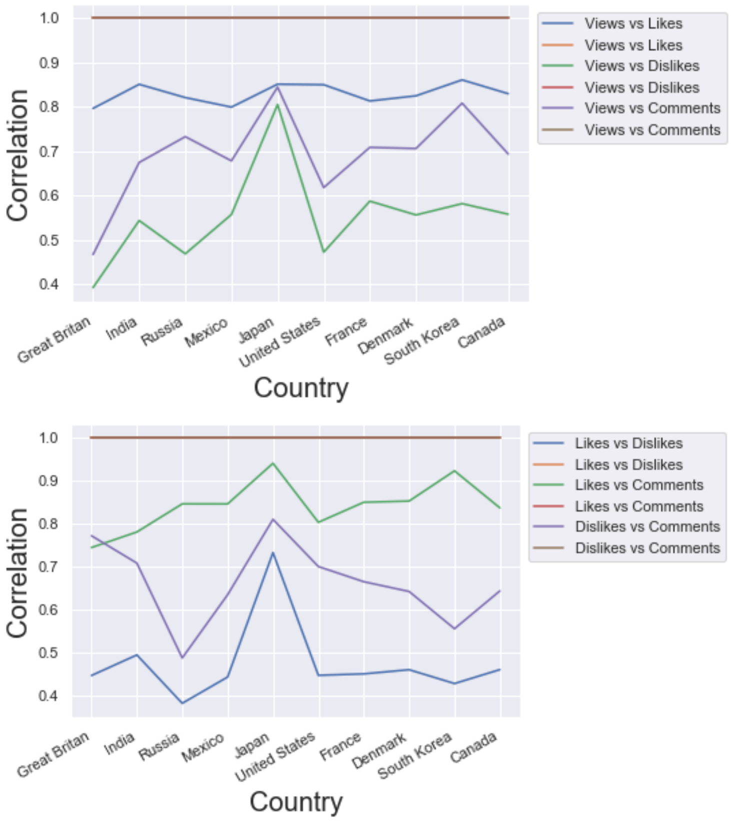

Fig. 8. Correlations between different attributes. The correlations are plotted on the y-axis and the countries are plotted on the x-axis.

5.1.3 comments_disabled, ratings_disabled

In this subsection, we analyse how user participation affects views. We compare the difference between the number of views when the comments are enabled or are disabled and when ratings are enabled or disabled is done.
Figure 9 shows the comparison of number of views on the trending videos when comments are enabled/disabled. From the graph,we observe that the videos for which comments are enabled get more views.

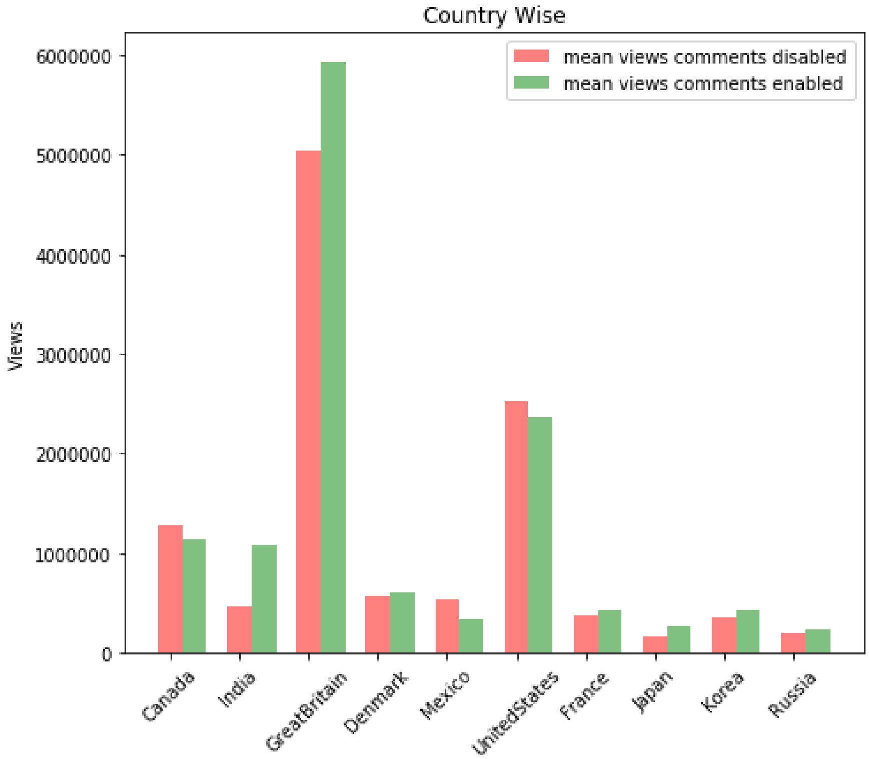

Fig. 9. Comparison of number of views when comments are enabled/disabled. Number of views is plotted on the y-axis and the countries are plotted on the x-axis. The videos for which comments are enabled get more views.

Figure 10 shows the comparison of number of views on the trending videos when ratings are enabled/disabled. From the graph,we can observe that the videos for which ratings are enabled get more views.
Videos with comments enabled and rating enabled yield more views which tells us that a trending video is a direct result of user participation.

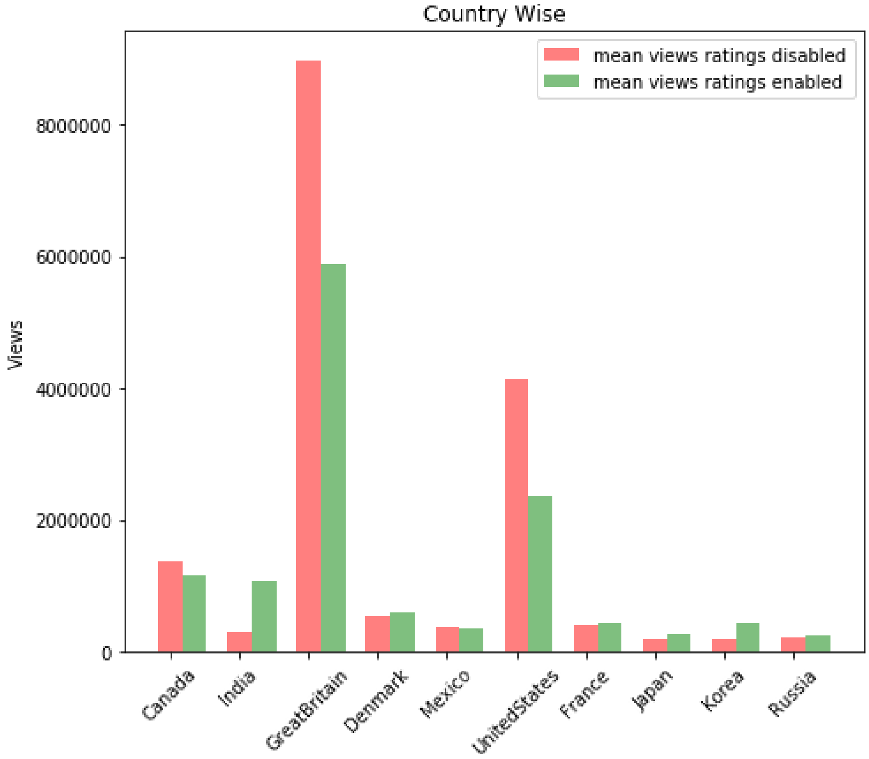

Fig. 10. Comparison of number of views when ratings are enabled/disabled. Number of views is plotted on the y-axis and the countries are plotted on the x-axis. The videos for which ratings are enabled get more views.

5.1.4 Publish Time

We analyse publish time to identify the time slot with the highest number of uploads, the time slot where highest and to see how publish time affects user participation and user engagement. We divide the time into four slots: (1) between 00:00 to 06:00, (2) between 06:00 to 12:00, (3) between 12:00 to 18:00, (4) between 18:00 to 24:00.

According to the analysis done in Figure 11, the most popular publish time slot is between 12:00 to 18:00 except in Japan, Mexico and Korea.One of the reasons for this could be because of the maximum involvement of users on the YouTube platform to take a break and re-energize in the middle of their day.

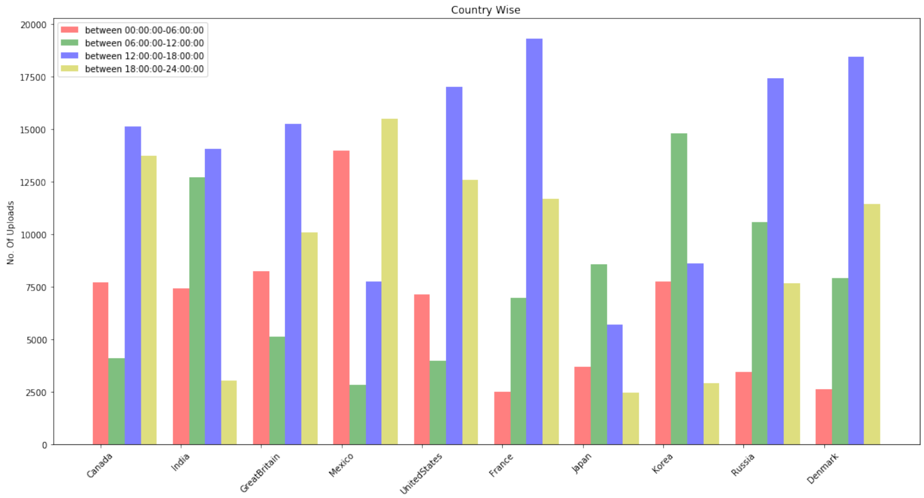

Fig. 11. Number of videos published in different time slots across different regions. Number of videos is plotted on the y-axis and the countries are plotted on the x-axis. The colored bars show the number of videos published in the different time slots

5.1.5 Categories popular in 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 12, we 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.

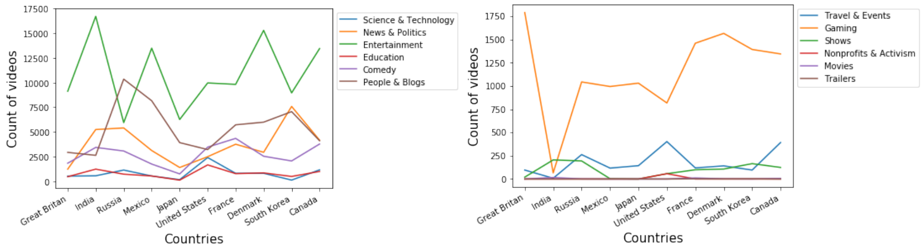

Fig. 12. Distribution of the trending videos in different categories. The count of the video is plotted on the y-axis and the country names are plotted on the x-axis

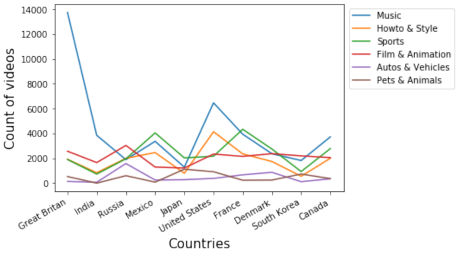

Fig. 13. Distribution of the trending videos in different categories. The count of the video is plotted on the y-axis and the country names are plotted on the x-axis

5.2 Prediction of numeric video attributes

In this study, we analyse phenomena and trends that arise on the video sharing platform, YouTube. For this purpose, we use a multiple linear regression approach.
We construct four prediction models with different target attributes and evaluate them on the basis of R-squared.

The regressors are defined as:
$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + ... + \beta_p x_{i,p}$
where, for i = 4:
Y_i = the dependent variable; $Y_i \in \{viewx, likes, dislikes, comments\}$
$x_{i,j}$ = the explanatory variables;
$x_{i,j} \in \{views, likes, dislikes, comments, time\_to\_trend, publish\_hour,$
$category\_id, tag\_count, comments\_disabled, ratings\_disabled, video\_error\}$
β_0 = y-intercept present(constant term)
β_p = the regression coefficients for each explanatory variable

Table 3. Values of Regression Coefficient (Weights).

Model No.	Target	x0	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
1	**Views**	-0.05	-0.0	-0.0	0.01	0.19	-0.08	0.01	-	0.52	1.01	-0.06
2	**Likes**	-0.07	0.0	0.0	-0.01	0.07	0.05	0.05	0.58	-	0.06	1.15
3	**Dislikes**	0.07	-0.0	0.0	-0.0	-0.04	0.03	0.0	1.10	0.06	-	0.50
4	**Comments**	0.03	0.0	-0.0	0.0	0.01	0.02	0.02	-0.07	1.20	0.53	-

X	Attribute Name
x0	category_id
x1	comments_disabled
x2	ratings_disabled
x3	video_error_or_removed
x4	timetotrend
x5	hour
x6	tag_counts
x7	log_views
x8	log_likes
x9	log_dislikes
x10	log_comments

The P-value for all attributes is less than 0.05.

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 four regressors and are therefore, excluded from further analysis.

5.2.1 Views Regressor

The global views regressor, that is, the model which does not take region as an input, is mathematically defined as:

$$Y = 11.59 - 0.05x_1 + 0.18x_2 - 0.08x_3 + 0.52x_4 + 1.00x_5 - 0.06x_6 \tag{2}$$

where,
Y = views
$X_i \in category\_id, time\_to\_trend, publish\_hour, likes, dislikes, comments$

The regression model reveals that views are most influenced by dislikes followed by likes. This implies that user participation has a high influence on user consumption.

We now consider videos specific to regions to analyse how regression coefficients vary across countries.

From Figure 14 , 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. It is also observed that for videos specific to Great Britain, likes influence views more than dislikes which is not in accordance with the global analysis.

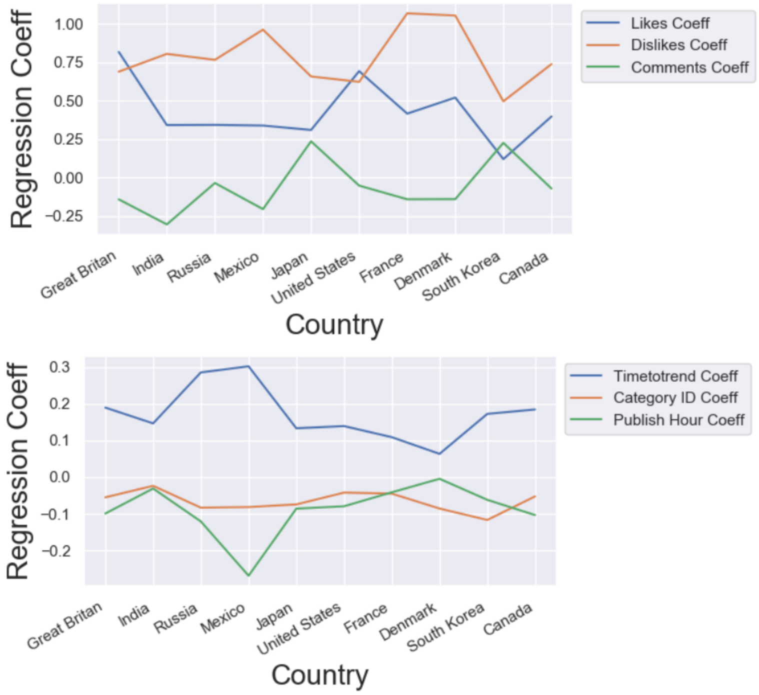

Fig. 14. Regression Coeff. for likes, dislikes, comments, time_to_trend, category_Id and publish_hour

France has the highest value for dislikes regression coefficient, hence number of dislikes has the largest influence on views in France. It is also observed that for videos specific to France, dislikes has the greatest regression coefficient in comparison to all other attributes which is in accordance with the global analysis.

Japan & South Korea have the highest value for comment count regression coefficient. This implies that Japan and South Korea have similar YouTube participation and consumption cultures which could be due to the actual cultural similarities that the two countries share. It is also observed that for videos specific to the two countries, dislikes has the greatest regression in comparison to all other attributes which is inline with the global analysis.

Mexico has the highest value for timeToTrend regression coefficient. This implies that videos specific to Mexico are likely to get more views in comparison to other countries for the same amount of days it takes for a video to get trending.

India has the highest value for category ID regression coefficient. This implies that genre of a video matters more to users specific to India in comparison to other countries.

Denmark has the highest value for publish hour regression coefficient. This implies that the time at which a video is published matters most in Denmark as compared to other countries and that Denmark specific videos are likely to get more likes than videos published at the same time in other countries.

5.2.2 Likes Regressor

The global likes regressor, that is, the model which does not take region as an input, is mathematically defined as:

$$Y = 7.66 - 0.07x_1 + 0.07x_2 + 0.04x_3 + 0.57x_4 + 0.05x_5 + 1.14x_6 \tag{3}$$

where,
Y = likes
$X_i \in category\_id, time\_to\_trend, publish\_hour, views, dislikes, comments$

The regression model reveals that likes are highly influenced by the number of comments followed by views. This implies that user participation prompts more user participation.

We now consider videos specific to regions to analyse how regression coefficients vary across countries.

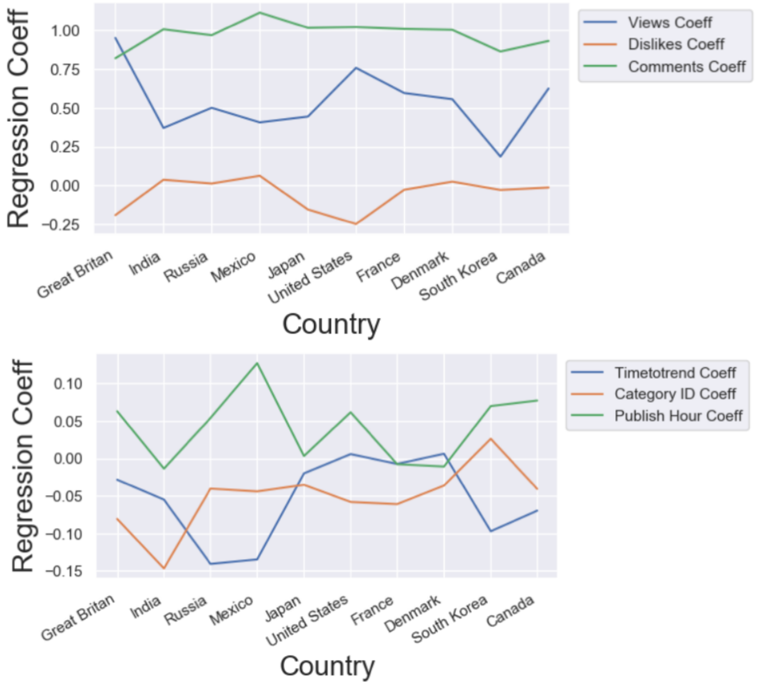

Fig. 15. Regression Coeff. for views, dislikes, comments, time_to_trend, category_Id and publish_hour

From Figure 15, 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. It is also observed that for videos specific to Great Britain, views influence likes more than comments which is not in accordance with the global analysis.
Mexico has the highest value for dislikes, comments and publish hour regression coefficients. This implies that user participation is likely to garner most likes in Mexico in comparison to other countries.
South Korea have the highest value for categoryId regression coefficient and hence, genre is plays an important role in generating likes for videos speific to South Korea.
United States has the highest value for the timeToTrend regression coefficient. This implies that videos specific to United States are likely to get more views in comparison to other countries for the same amount of days it takes for a video to get trending.

5.2.3 Dislikes Regressor

The global dislikes regressor, that is, the model which does not take region as an input, is mathematically defined as:

$$Y = 4.65 + 0.07x_1 - 0.03x_2 + 0.02x_3 + 1.09x_4 + 0.05x_5 + 0.494x_6 \qquad (4)$$

where,
Y = likes
$X_i \in category\_id, time\_to\_trend, publish\_hour, views, likes, comments$

The regression model reveals that dislikes are highly influenced by the views followed by comments. This implies that videos which more views and comments are less likely to receive backlash.

We now consider videos specific to regions to analyse how regression coefficients vary across countries.

From Figure 16, it is evident that Mexico has the highest value for likes regression coefficient, hence number of likes has the largest influence on dislikes in Mexico. It is also observed that for videos specific to Mexico, views influence dislikes which is in accordance with the global analysis.

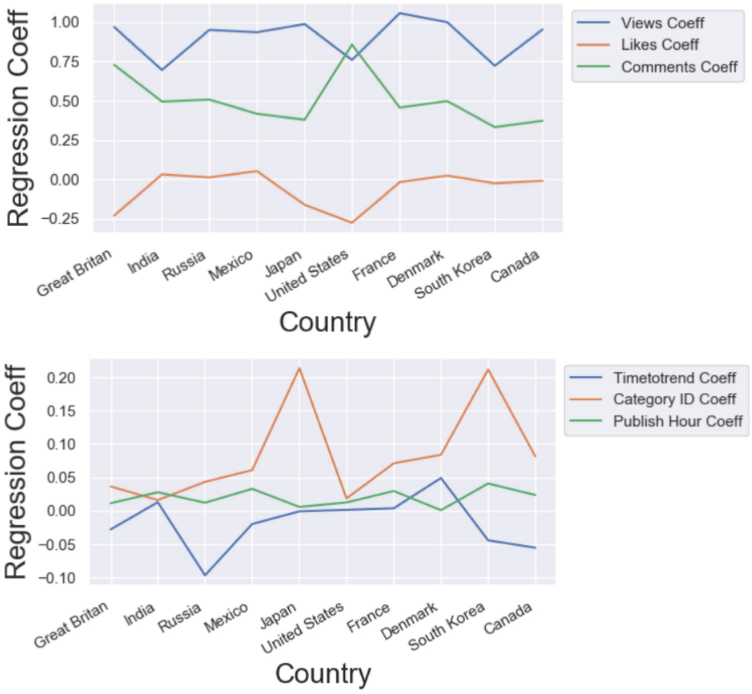

Fig. 16. Regression Coeff. for views, likes, comments, time_to_trend, category_Id and publish_hour

France has the highest value for views regression coefficient, hence number of views has the largest influence on dislikes in France. It is also observed that for videos specific to France, views has the greatest regression coefficient in comparison to all other attributes which is in accordance with the global analysis.
United States has the highest comments regression coefficient. This implies that people in the United States are more likely to derive their dislikings from other people's opinions of a video.
Japan & South Korea have the highest value for categoryId regression coefficient. This implies that Japan and South Korea have similar YouTube participation and consumption cultures which could be due to the actual cultural similarities that the two countries share. It is also observed that for videos specific to the two countries, views has the greatest regression coefficient in comparison to all other attributes which is in accordance with the global analysis.
Denmark has the highest value for timeToTrend regression coefficient. This implies that videos specific to Denmark are likely to get more dislikes in comparison to other countries for the same amount of days it takes for a video to get trending.
South Korea has the highest value for publish hour regression coefficient. This implies that the time at which a video is published matters most in South Korea

as compared to other countries and that South Korea specific videos are likely to get more dislikes than videos published at the same time in other countries.

5.2.4 Comments Regressor

The global comments regressor, that is, the model which does not take region as an input, is mathematically defined as:

$$Y = 5.65 + 0.03x_1 + 0.005x_2 + 0.01x_3 - 0.07x_4 + 1.2x_5 + 0.52x_6 \tag{5}$$

where,
Y = likes
$X_i \in category\_id, time\_to\_trend, publish\_hour, views, likes, dislikes$

The regression model reveals that comments are highly influenced by the likes followed by dislikes. This implies ratings prompt users to express their opinions of a video. We now consider videos specific to regions to analyse how regression coefficients vary across countries.

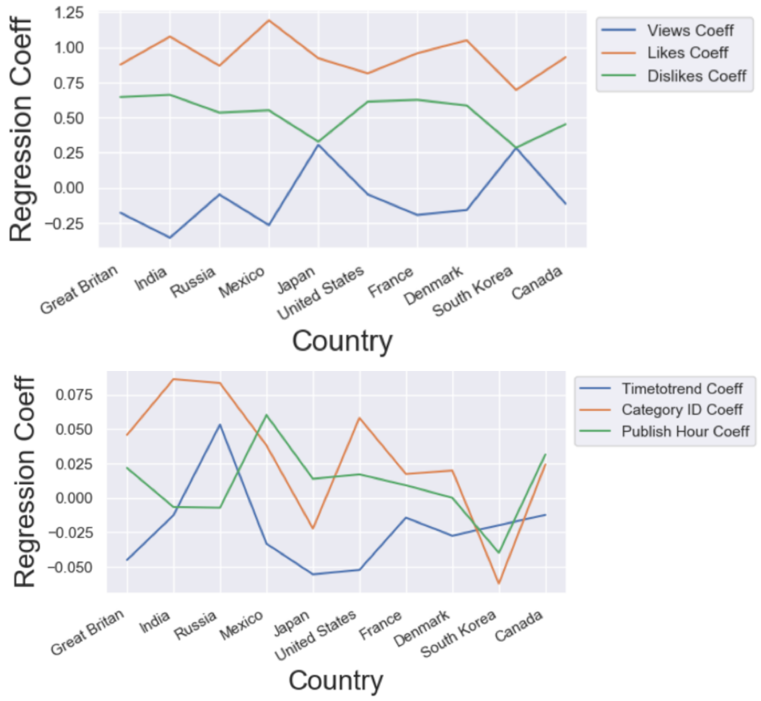

Fig. 17. Regression Coeff. for views, likes, dislikes, time_to_trend, category_Id and publish_hour

From Figure 17, it is evident that Japan & South Korea have the highest value for the views regression coefficient. This implies that Japan and South Korea have similar YouTube participation and consumption cultures which could be due to the actual cultural similarities that the two countries share. It is also observed that for videos specific to the two countries, likes has the greatest regression coefficient in comparison to all other attributes which is in accordance with the global analysis.
Mexico has the highest value for Likes and publishTime regression coefficients.
India has the highest value for dislikes regression coefficient. This implies that the more backlash a video receives, the more likely are users to express their opinion of the video in India as compared to other countries.
Russia has the highest value for the timeToTrend regression coefficient.
India has the highest value for categoryId regression coefficient closely followed by Russia.
Denmark has the highest value for likes regression coefficient.

5.2.5 Evaluation of Regressors

We evaluate the four prediction models constructed on the basis of R-squared.
R-squared measures the strength of the relationship that exists between our regression model and the explanatory variables.
The higher the value of R-squared, the better is the model as a higher R-squared implies that there are smaller differences between the fitted values and the observed data.

We compare the R-squared of the four regressors that have been formulated.

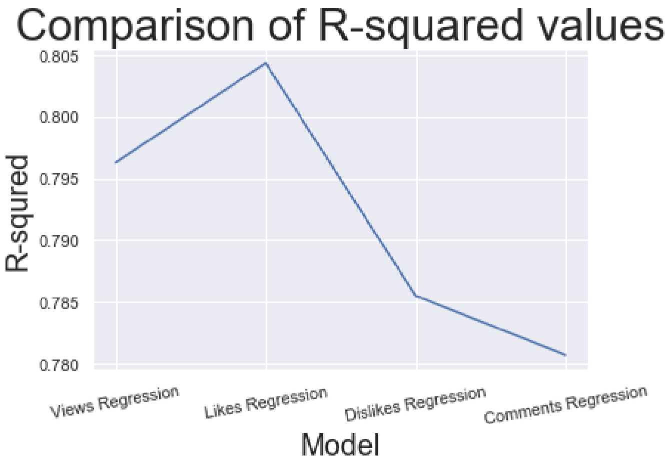

Fig. 18. describes the R-squared values for our four defined regression models.

It is evident from Figure 18 that likes regressor has the highest R-squared and hence is the best fitted and most accurate model. This also implies that there is a greater dependence between likes and other attributes in comparison. Comments regressor has the least value of R-squared which means that whether a video will be receiving more or less comments in least dependent on other attributes of the video.
The sharp fall in R-squared from likes regressor to dislikes regressor is also worth noting.

5.3 Classification of YouTube video titles based on region

In this study we have two text based descriptive, that is, categorical features: tags and titles.

Tags are descriptive keywords that are added to videos to help viewers find content close to their search. Tags can be used by YouTube channels to increase their user engagement.

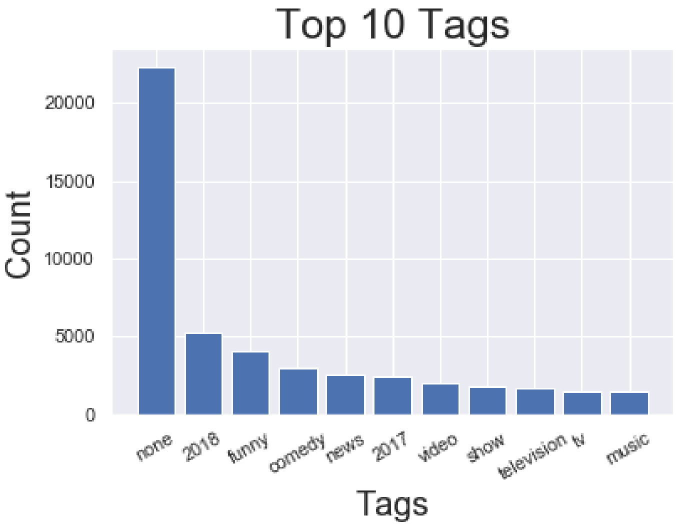

Fig. 19. Top 10 recurring tags. The top 10 tags are plotted on the x-axis and their count is plotted on the y-axis.

Figure 19 describes the top 10 tags used in trending videos. We observe from this analysis that majority of the trending videos don't have any tags. Tags might increase user engagement but are not the only important factor to make a video trending.

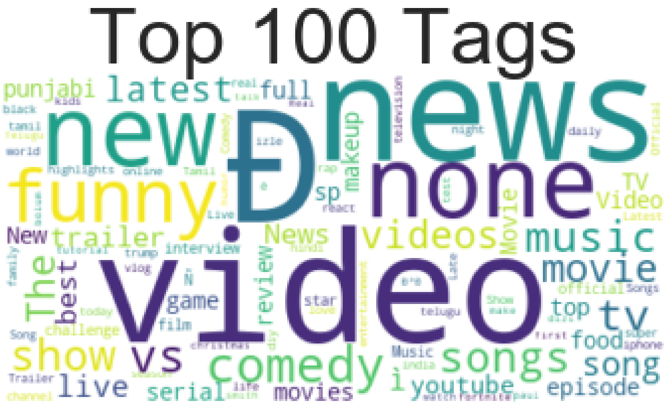

Fig. 20. Word cloud representing the top 100 most recurring tags in YouTube videos

Figure 20 describes the top 100 most recurring tags which are used in the trending YouTube videos.

Further, we perform text classification for title feature of the data set. Title describes the title description of the YouTube video.
For text classification, we use region feature as our classes on the basis of which we classify the YouTube video titles.

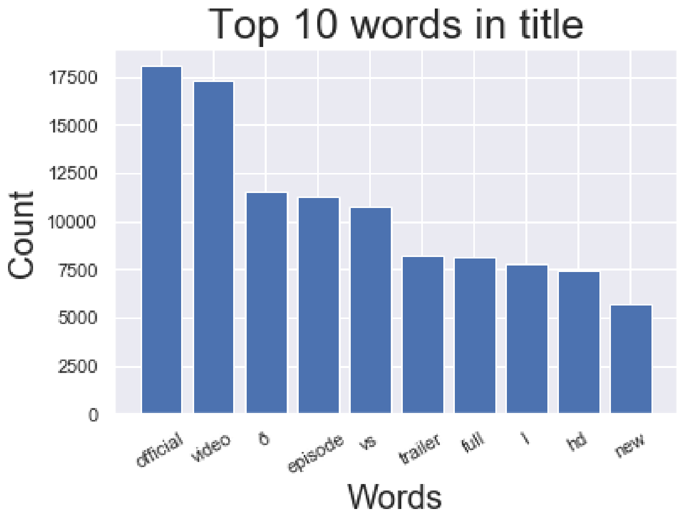

Fig. 21. Top 10 recurring words in title. The top 10 words found in title is plotted on the x-axis and their count is plotted on the y-axis.

Titles are an important factor for a user to consider whether to watch a video or not as it describes the content of the video. Figure 21 describes the top 10 recurring words used in trending video titles. We observe that the most common word found in titles is ’official’ with count approximately equal to 24,000. This implies that users are more willing to engage in content which seems verified and coming from original sources. ’official’ is followed by ’video’.

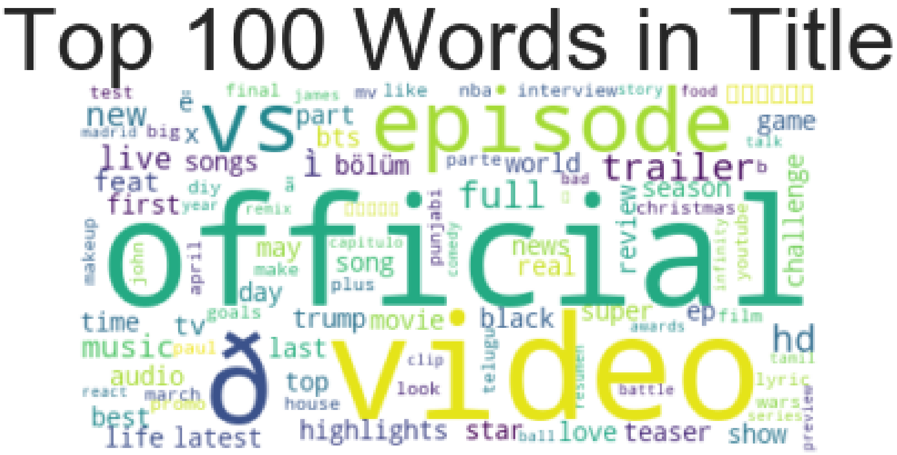

Fig. 22. Word cloud representing the top 100 words used in the trending videos

Figure 22 describes the top 100 most recurring words which were are used in titles of the trending YouTube videos.

We test two models for feature extraction namely, BOW and TF-IDF, for five configurations of training and testing split. Both models are compared on the basis of MultinomialNB Accuracy, Precision and Recall for the five configurations.

5.3.1 BOW model implemented on Title

BOW model uses frequency of a word which is represented using a vector for feature extraction. It calculates the frequency across all documents.

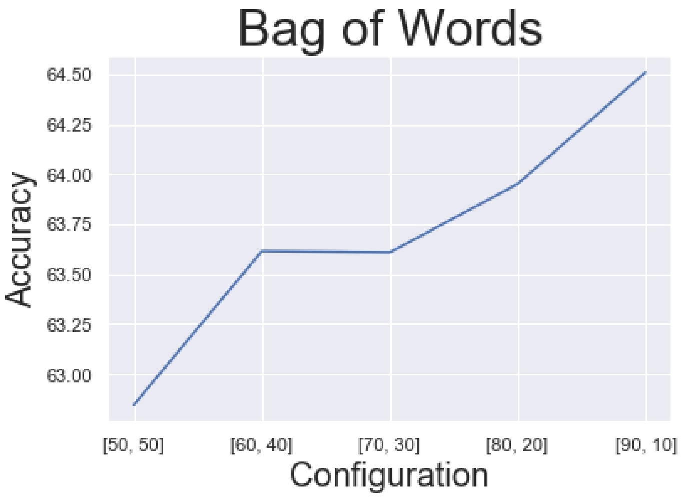

Fig. 23. BOW Model: Different configurations of training and testing set are plotted on the x-axis and the accuracy obtained for each set is plotted on the y-axis.

From Figure 23, we observe that the Text Classification model using BOW is most accurate for 90:10 configuration.

5.3.2 TF-IDF Model implemented on Title

In Term Frequency(TF), the model takes the number of words occurred in each document.
But the issue with Term Frequency is that more weight is given to longer documents.
While IDF(Inverse Document Frequency) on the other hand is used to measure the amount of information provided by a given word across the document.
Combination of these two: TF-IDF(Term Frequency-Inverse Document Frequency) is used to normalize the document term matrix. TF-IDF is a combinations of both TF and IDF.

We take five configurations of the training and testing set on the basis of MultinomialNB Accuracy.

Figure 24 shows that the Text Classification model using TF-IDF is most accurate for 90:10 configuration.
Words in a document that have a high TF-IDF are the most frequently occurred words in the given documents and these words must not be present in the other documents. Hence, these words that have a high TF-IDF must be signature words.

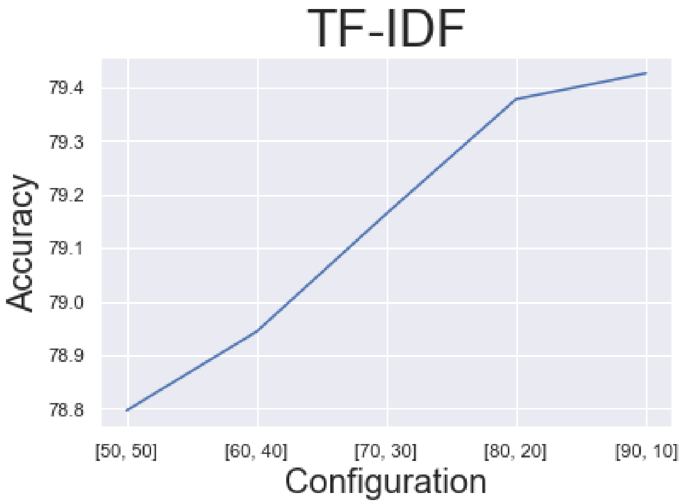

Fig. 24. TF-IDF Model. Different configurations of training and testing set is plotted on the x-axis and the accuracy obtained for each set is plotted on the y-axis.

5.3.3 Comparison of BOW and TF-IDF

We compare Text Classification Model using BOW and TF-IDF for different configurations of training and testing split against MultinomialNB Accuracy.

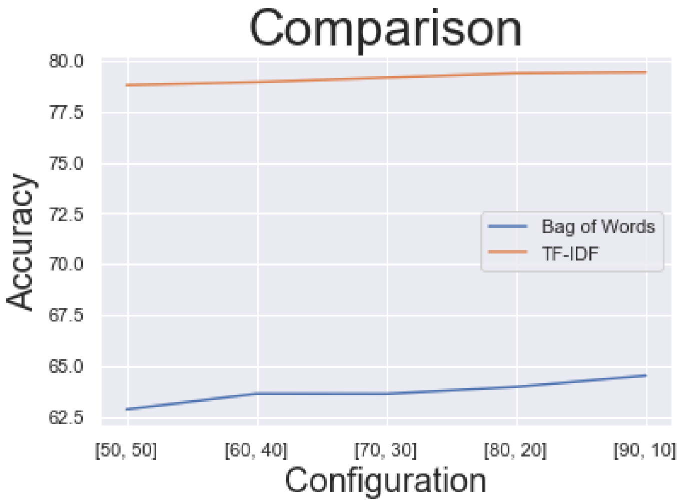

Fig. 25. BOW Vs. TF-IDF. For different configurations of training and testing data(plotted on x-axis), we plot accuracy for each on the x-axis for comparison. We observe that TF-IDF model is more accurate than BoW model.

From Figure 25, we observe that TF-IDF is more accurate than BOW. This is becuase TF-IDF normalises the document matrix.

We now evaluate our classifier using precision and recall metrics.
Precision refers to the percentage of results which are relevant and recall refers to the percentage of the total relevant results which have been correctly classified by the model.

Table 4. Precision , Recall Values for BOW and TF-IDF at 90:10 Split.

Model	Class	Instances	Precision(%)	Recall(%)
BoW	Great Britan	38,916	25.8306	47.4059
	India	37,352	89.1669	66.4376
	Russia	40,739	86.8482	74.1712
	Mexico	40,451	59.8052	80.5906
	Japan	20,523	86.0290	85.9828
	United States	40,949	85.3354	26.8532
	France	40,724	94.2101	89.4972
	Denmark	40,840	59.7226	43.6912
	South Korea	34,567	61.9147	42.4476
	Canada	40,881	59.3632	68.4871
TF-IDF	Great Britain	38,916	49.5194	48.2384
	India	37,352	92.3124	73.5004
	Russia	40,739	86.6273	81.8777
	Mexico	40,451	62.6614	79.9301
	Japan	20,523	90.1453	90.8548
	United States	40,949	99.1299	87.2666
	France	40,724	94.8601	90.7053
	Denmark	40,840	98.5043	82.1673
	South Korea	34,567	96.5996	91.9929
	Canada	40,881	57.2845	75.0601

From Table 4, we can see that TF-IDF gives a higher precision and recall value than BOW which further emphasises on how TF-IDF is more accurate than BOW.
For both BOW and TF-IDF models, we see that for class Great Britain the value of precision and recall are the least which implies that the probability of correctly classifying this class is the least in comparison to all other classes.
Given the high values of precision and recall for TF-IDF model, we can say that our model is fairly accurate and good at classifying video titles.

5.4 Some Global Analysis

Global Analysis of the trending videos are done here. Factors like Upload Time, Likes, Dislikes, Comment Count, Trending time varies from video to video. In this subsection, we study the attributes globally for the trending videos.

5.4.1 Analysis of attributes: number of views, number of likes, number of dislikes, number of comments, time taken to trend

In this subsection, we compile the results from analysis of a trending YouTube video globally. We visualize analysis of the attributes vs number of videos associated with the value of attributes. From Figure 26, we note that most of the trending videos have views lesser than or equal to 1 million. Further, the majority of trending videos have 50,000 likes or less(Figure 27) and 20,000 dislikes or less (Figure 28). The comment count in majority of the trending videos is 4,000 or less (Figure 29). 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 30).

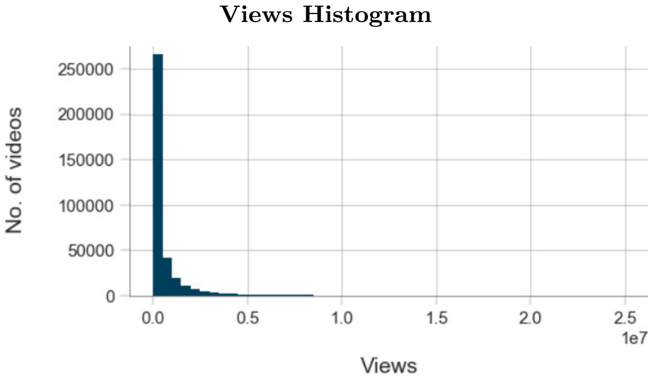

Fig. 26. Number of views associated with videos. The number of videos is plotted on the y-axis and the 1e7 of views associated is plotted on the x-axis

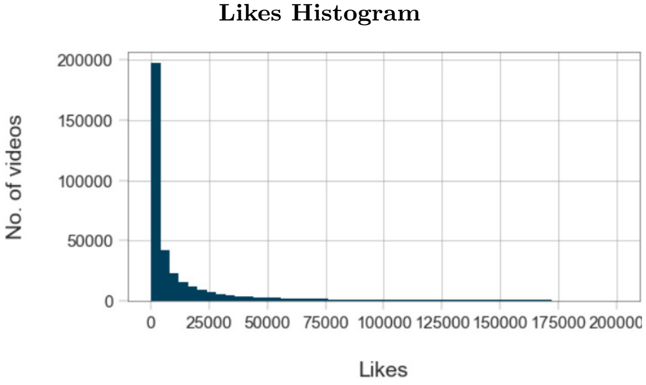

Fig. 27. Number of likes associated with videos. The number of videos is plotted on the y-axis and the number of likes associated is plotted on the x-axis

Dislikes Histogram

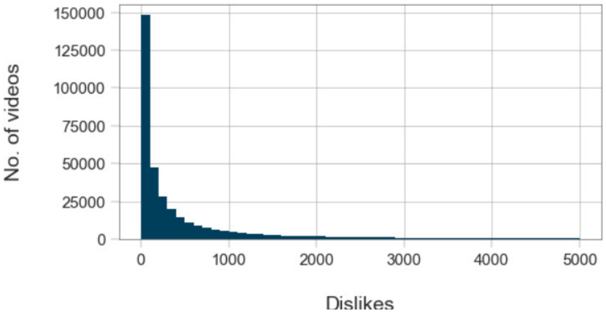

Fig. 28. Number of dislikes associated with videos. The number of videos is plotted on the y-axis and the number of dislike associated is plotted on the x-axis

Comments Histogram

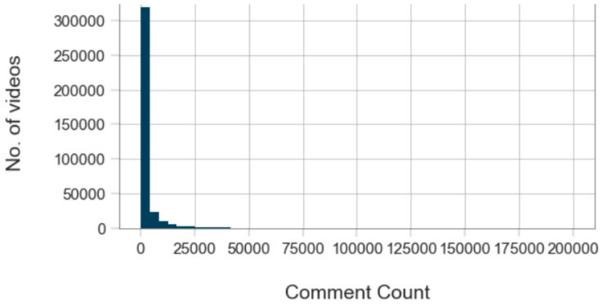

Fig. 29. Comment count associated with videos. The number of videos is plotted on the y-axis and the number of comments associated is plotted on the x-axis

Time taken to trend Histogram

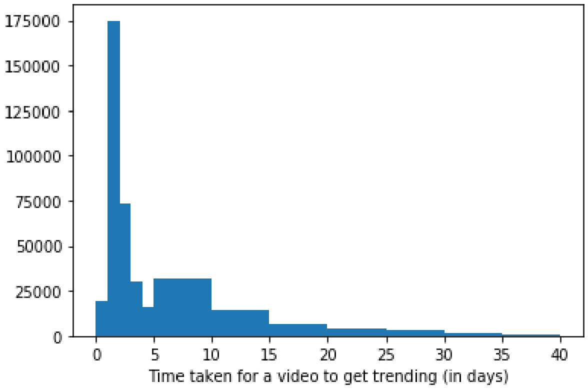

Fig. 30. Time taken by videos to trend. The number of videos is plotted on the y-axis and the time take for the video to trend is plotted on the x-axis

5.4.2 Correlations between the attributes

From Figure 31, we observe that there is a strong positive correlation between the number of likes and the number of comments of trending videos. There is also a strong positive correlation between the number of views and the number of likes. However, there is a slightly weaker correlation between the number of dislikes and the comment count. Neither there is a strong correlation between timetotrend and the other attributes. We observe that timetotrend is the most related to the number of views compared to the other attributes.

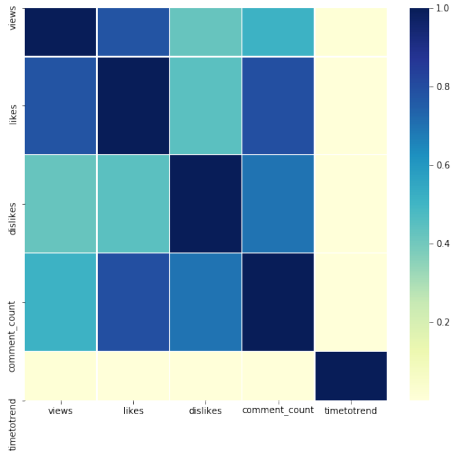

Fig. 31. Correlation matrix between different attributes

5.4.3 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
We analyse how the number of views and the time taken to trend vary in the given cases.

Views Vs. comments disabled and ratings disabled

Figure 32 tells us about the number of videos which made it to the list of trending videos in the different cases of enabling/disabling of ratings and comments. While Figure 33 gives us the mean views in each case.
After studying the graphs and data set, we conclude that since number of in-

stances 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 34 and Figure 35 are being used to represent the same.
After studying the data, we conclude that: Ratio of total number of views on videos with both comments and ratings disabled to total number of views received by videos that have 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 maximum number of views on videos with both comments and ratings disabled to maximum number of views on videos with both ratings and comments enabled = 0.1468378050584913 This implies that, maximum number of views on videos with both comments and ratings disabled is less than maximum number of views on videos with both comments and ratings enabled.

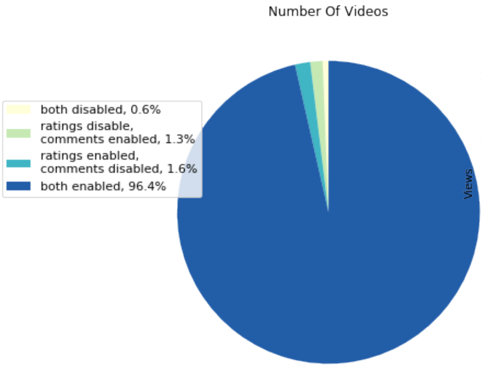

Fig. 32. Number of Videos. The different colored parts show the four case represented in the legend

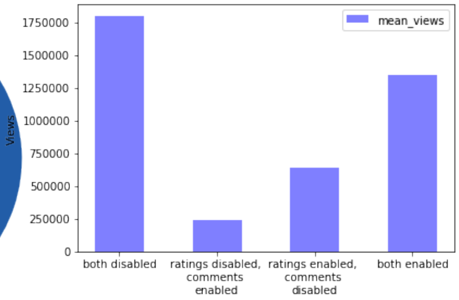

Fig. 33. Mean Views for each case. Mean views is plotted on the y-axis while the case are represented on the x-axis

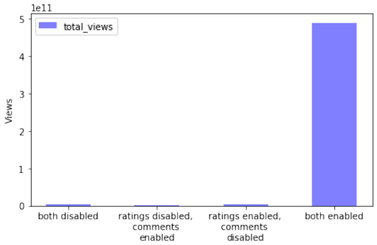

Fig. 34. Total Views associated with each case

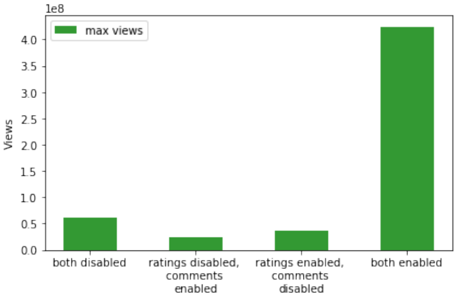

Fig. 35. Max Views associated with each case

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 36 depicts the same.

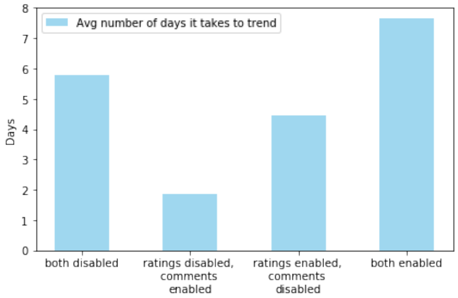

Fig. 36. Average number of days taken by a video to get trending. The number of days taken for the video to get trending is plotted on the y-axis for the different cases represented on the x-axis

5.4.4 Analysis of the Title length Analysis between title length and number of views.

We plot a scatter plot in order to examine the relationship between title length and number of views. 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 37), we observe that the videos with views over 100 million have titles of length between 33 and 65 characters approximately.

By looking at the scatter plot, we observe that there is no relationship between the length of the title and the number of views received by that video.

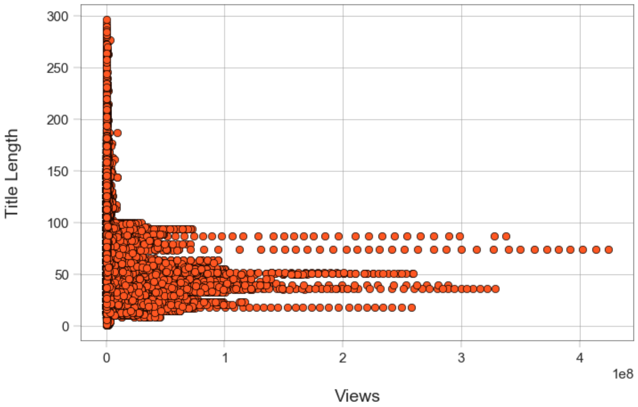

Fig. 37. Number of views vs Title Length

5.4.5 Analysis of publish time

Analysis of upload time based on publish time

Figure 38 tells us about the number of videos uploaded in different time slots. We note that majority of the videos are uploaded between 12:00-18:00.

Analysis of likes based on publish time

Figure 39 tells us about the number of likes on videos which are uploaded in different time slots. We note that a video gets maximum likes when was published in the time slot of 00:00-06:00

Analysis of dislikes based on publish time

Figure 40 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 based on publish time

Figure 41 tells us about the number of comments on videos which are uploaded in different time slots. We note that a video gets maximum comments when published in the time slot of 00:00-06:00

Upload Time

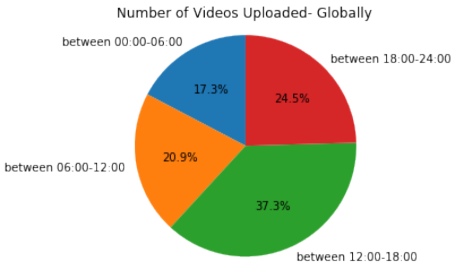

Fig. 38. Upload Time of videos.

Mean Likes

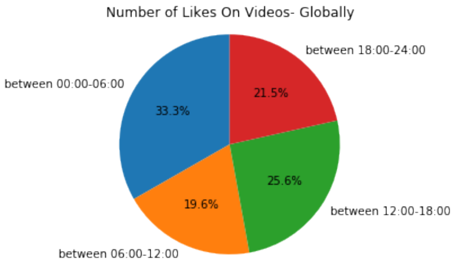

Fig. 39. Likes depending on the publish time.

Mean Dislikes

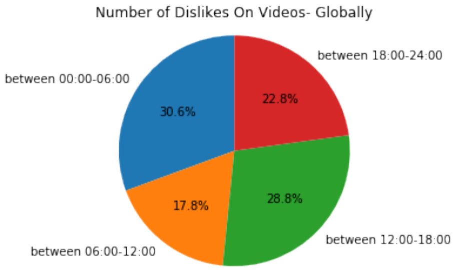

Fig. 40. Dislikes depending on the publish time

Mean Comments

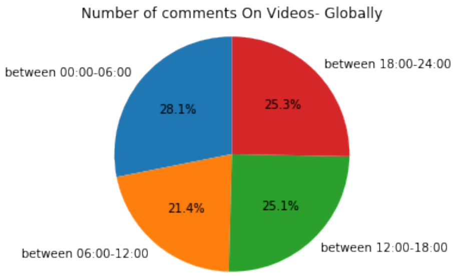

Fig. 41. Comment count depending on the publish time

5.4.6 Time taken for a video to trend in different categories

From Figure 42 , We conclude that the time taken for a video to trend highly depends on the category to which that video belongs. We can infer that it takes maximum no. of days(i.e 18 days) for a video of category 'how to style' to trend whereas it takes least number of days (i.e 2 days) for a video of category 'shows' and 'Non-profits and activism' to trend.

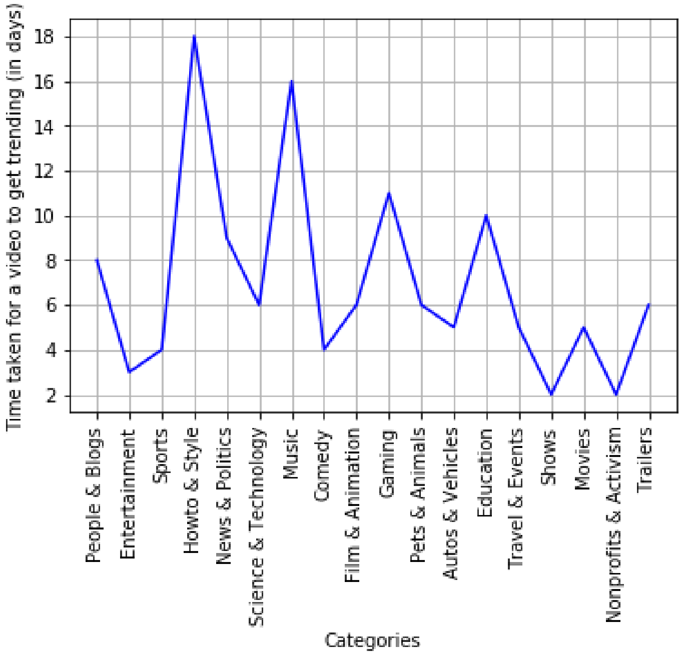

Fig. 42. Average time taken(in days) for a video to get trending in different categories. The time taken is represented on the y-axis corresponding to the categories which are plotted on the x-axis

6 Future Work and Conclusion

6.1 Conclusion

While web platforms such as YouTube provide access to diverse cultural products, our study reveals that consumption of content is rather constrained by cultural differences. The result indicates that user participation has high dependence on other aspects of a video such as publish time, category etc.
This paper creates value for content creators and advertisers on the platform as the analysis conducted provides insights into how one can increase user engagement and reach of their videos which ultimately leads to profit maximisation.
With the dawn of the social media age, a new concept of ”internet trolls” has also emerged. Internet trolls target content creators and create extreme backlash. Many creators have claimed that this negatively impacts their mental health, which is why many a times they choose to disable ratings and comments. Our study creates value for such creators as it enables them to navigate the platform space and increase profits through investment in other aspects of a video.
Morever, in time such as the current scenario, that is the COVID19 pandemic,

our study can also be used to push public service announcements. This study however, lacks in identifying whether a video is positively or negatively received.

6.2 Future Work

For future work, Sentiment analysis or opinion mining can be done,in which we can analyze sentiments, evaluations, opinions, attitudes, and emotions of users as their opinions or sentiments or attitude are being expressed on the videos that they watch.

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.

Our study can also be expanded to understand how social situations such as the current worldwide lockdown in wake of the COVID19 pandemic, impacts content consumption.

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. Acar, A.: Antecedents and consequences of online social networking behavior: The case of facebook. Journal of Website Promotion **3**(1-2), 62–83 (2008). https://doi.org/10.1080/15533610802052654, `https://doi.org/10.1080/15533610802052654`
2. Bärtl, M.: Youtube channels, uploads and views: A statistical analysis of the past 10 years. Convergence **24**(1), 16–32 (2018). https://doi.org/10.1177/1354856517736979, `https://doi.org/10.1177/1354856517736979`
3. Cheng, X., Dale, C., Liu, J.: Statistics and social network of youtube videos. In: 2008 16th Interntional Workshop on Quality of Service. pp. 229–238 (2008)
4. Khan, M.L.: Social media engagement: What motivates user participation and consumption on youtube? Computers in Human Behavior **66**, 236 – 247 (2017). https://doi.org/https://doi.org/10.1016/j.chb.2016.09.024, `http://www.sciencedirect.com/science/article/pii/S0747563216306513`
5. McRoberts, S., Bonsignore, E., Peyton, T., Yarosh, S.: "do it for the viewers!" audience engagement behaviors of young youtubers. In: Proceedings of IDC 2016 - The 15th International Conference on Interaction Design and Children. pp. 334–343. Proceedings of IDC 2016 - The 15th International Conference on Interaction Design and Children, Association for Computing Machinery, Inc (Jun 2016). https://doi.org/10.1145/2930674.2930676, 15th International Conference on Interaction Design and Children, IDC 2016 ; Conference date: 21-06-2016 Through 24-06-2016
6. Xu, W.W., Park, J.Y., Kim, J.Y., Park, H.W.: Networked cultural diffusion and creation on youtube: An analysis of youtube memes. Journal of Broadcasting & Electronic Media **60**(1), 104–122 (2016).

https://doi.org/10.1080/08838151.2015.1127241, https://doi.org/10.1080/08838151.2015.1127241