YOUTUBE STATISTICAL ANALYSIS

Group 11:

Laxmi Shiva Prasad

Manavi Reddy Vemula

Navaneeth Reddy Chinti reddy

Tilak Reddy Toom

Vivek Kothuru

PROJECT INTRODUCTION

Problem

Many youtubers are having problems with their content because it fails to meet the expectations of their viewers. On YouTube, there are numerous domains/categories, and while some youtubers release videos based on trends or different content to reach out to the public. But this does not always work. Youtubers face the problem of not knowing when to post their videos in order to reach a large audience. On the other side, users have difficulty locating the video they want to watch in a specific category.

About Data

The dataset contains data on daily trending YouTube videos for several months. Data is available for the US, GB, DE, CA, FR, and IN (the United States, Great Britain, Germany, Canada, France and India respectively), with up to 200 trending videos listed each day.

The data for each region is stored in a separate file. The video title, channel title, publish time, tags, views, likes, dislikes, description, and comment count are all included in the data. A category id field is also included in the data, which varies by region. Locate the associated JSON to retrieve the categories for a specific video. Each of the dataset's six regions has one of these files.

CHANGES SINCE THE PROPOSAL

We have decided to add a sentimental analysis to determine the polarity of comments in various categorical YouTube videos. This will provide us with additional information about how users are reacting to specific categories. In addition, six countries' data will be subjected to various types of cumulative analysis.

DATA

We discovered the data on the kaggle website; the data was perfectly organized, with each region's data in its own file. The video title, channel title, publish time, tags, views, likes, dislikes, description, and comment count are all included in the data.

After inspecting the data, we used pandas methods (preprocessing techniques) to put the correct timestamp for different columns in order to keep the timestamps accurate for different columns (publish time, trending date, trending year, etc.). Further, we removed the NAN values from the description column and replaced them with the empty string. We loaded the dataset for further proceeding with our project after finishing the preprocessing techniques.

EXPLORATORY DATA ANALYSIS

Our main focus is to perform different EDA on the data. We decided to analyze the data in the below various forms:

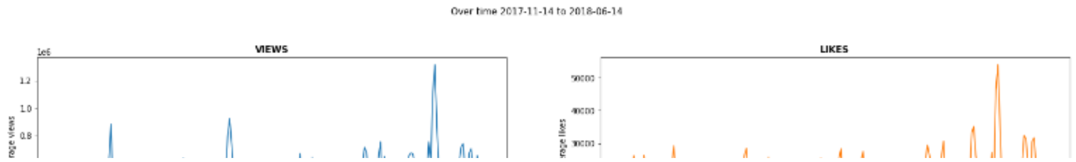
- Correlation Analysis
- EDA w.r.t to Time Series
- Most VIEWED or LIKED or DISLIKED video
- EDA related to channels
- Top 10 most trended channels
- Function to gather channel_title and its data per video
- Get the statistics for the given channels
- EDA on category attribute
- Analysis of cumulative data
- Sentimental and text analysis to find polarity for different categories

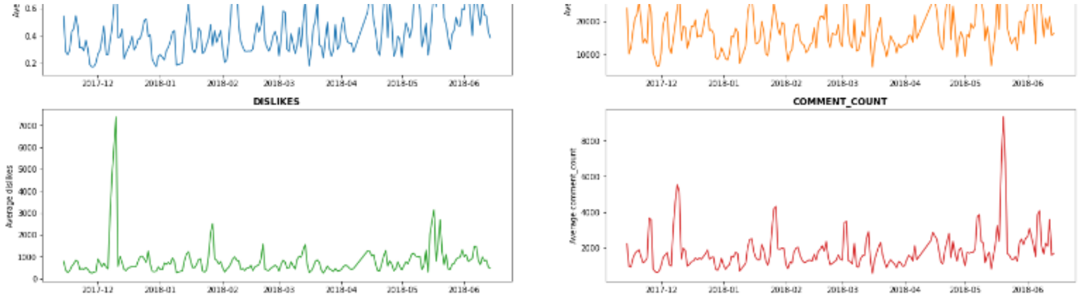
We have implemented until the beginning of the EDA w.r.t. time series. We discovered a significant relationship between views and likes for the France country data in various categories, with views and likes being highly correlated. We have come up with a graph of the views, dislikes, and likes. This demonstrates that feedback can help the creator understand the analysis of his video content's quality.

VISUALIZATION

Data for time series consists of timestamp columns corresponding to the video's release date, views, likes, dislikes, and comment count. We want to look at how frequently different categories of videos are released per week, month, and year, as well as how rapidly subscribers are increasing for their channel growth based on their content. This allows creators to see how far they've come in their youtube career and how well they're doing in terms of providing valuable content to the users. Below are the few visualizations drawn. We can find some useful insights from removed stats, total_videos_per_year.

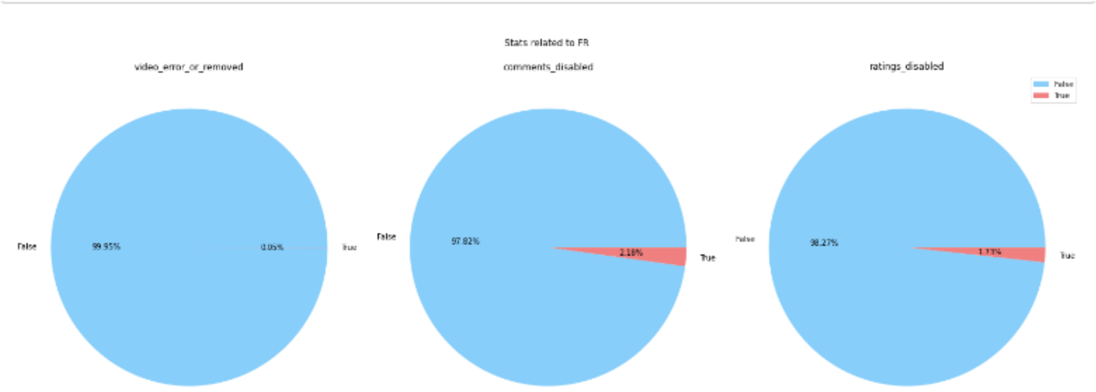
Outcomes





The above graph tells us about the average plot time period for likes, views, dislikes and

Removed stats

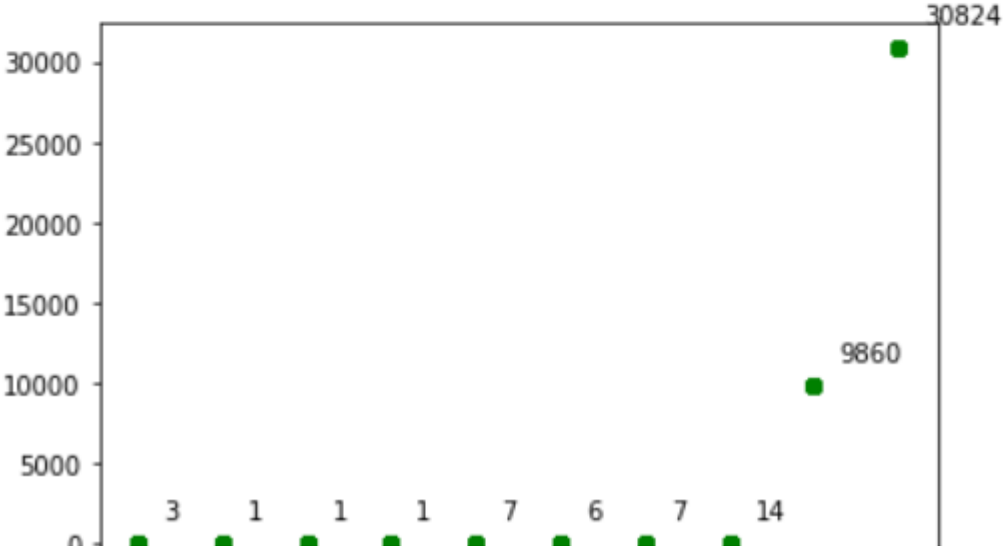


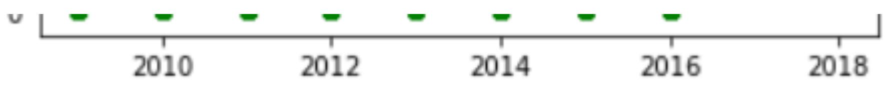
Using removed stats, we can draw the insights for video_error. It says that 99.95% of videos are not removed and 0.05% are removed because of streaming play error.

For comments disabled distribution, we can say that 2.18% video comments are disabled and the remaining 97.82% are not disbaled.

For rating disabled distribution, we can say that 1.73% of video ratings are disabled and the remaining are not disabled.

total_videos_per_year





This plot tells us about the total number of videos released in Great Britain from (2010-2018).

ML ANALYSIS

We first tried using a decision tree algorithm to train the dataset to predict the title based description. For this algorithm, we got around 70% accuracy. After experimenting with various ML algorithms, we came up with the RandomForestClassifier, which achieved 72% accuracy. Upon finding an optimal algorithm, we came across the use of ensemble learning research paper and applied the combined algorithms of (Random Forest Classifier, Decision Tree Classifier, XGB Classifier, KNeighborsClassifier, MultinomialNB) which achieve 98% accuracy. With this efficiency we were able to predict the maximum correct labels of titles based on description for different categories.

REFLECTION

- What is the most challenging part of the project that you’ve encountered so far?

The most challenging part of the project that we’ve encountered so far is finding the time series visualization for different countries and merging them together to draw the insights for numbers of videos published in respective months.

- What are your initial insights?

As we checked the data, we initially thought that the USA has the better contribution to the youtube platform compared to the remaining countries because it has more categories and the number of videos released per month were more than other countries.

- Are there any concrete results you can show at this point? If not, why not?

Till today, we have come up with results that are obtained by implementing the EDA related to the channels w.r.t the time series.

- Going forward, what are the current biggest problems you’re facing?

Going forward, the problem we are facing is implementing sentimental analysis to find the polarity of different categories.

- Do you think you are on track with your project? If not, what parts do you need to dedicate more time to?

We are on track with our project as scheduled before, finishing the sub-parts of the project depending on our teammates' expertise in specific areas.

- Given your initial exploration of the data, is it worth proceeding with your project, why? If not, how will you move forward (method, data etc)?

Yes, it is worth continuing with the project because we have enough data to perform all of the methods mentioned in the initial presentation. Based on our results so far, the data was sufficient to train and test the model, as well as achieved the desired results . If in case we need more data to get the even more optimal solutions we're ready to use Youtube scrapper APIs in addition to the current data.

NEXT STEP

We are left with 15 days until 28th November. We have completed 3/4 of the project and we are left with cumulative analysis, sentimental and text analysis. For the remaining 15 days, we will divide these two pieces of work among ourselves. Few of the team members will proceed with implementing the EDA methods to visualize all the different category videos for the mentioned six countries. On the other hand, other members will be working on sentimental analysis to produce polarity of categories on YouTube videos and comments polarity to find (positive, negative, neutral) for a particular video. Before 28th November, we’ll accomplish all the tasks in the project and our team will be ready for presentations :).