Analysis of the YouTube Trending Dataset

*Note: Sub-titles are not captured in Xplore and should not be used

1st Sheel Shah
Electrical Engineering department
IIT Bombay
Mumbai, India
sheelfshah@gmail.com

2nd Sumeet Kumar Mishra
Electrical Engineering department
IIT Bombay
Mumbai, India
sumeet@iitb.ac.in

***Abstract*—YouTube is by far the largest online video sharing platform, and it continues to grow rapidly. The platform is perfect for mass advertising to a certain target audience and hence understanding several aspects about YouTube is instrumental for those who want to advertise as well as those creators who want ad revenue on their videos. Over 500 hours of videos are uploaded on YouTube every minute, that is 720,000 hours each day. For every location, YouTube classifies videos as 'trending' and these videos are shown to everyone in the region, hence boosting the performance of these videos. Hence, a model that can categorize trending videos is highly useful for a firm that wants to advertise to a particular audience through a highly viewed video. In this paper we have presented the report of our analysis of the videos that make it to the trending charts. We have used the dataset available on Kaggle for our analysis.**

I. Introduction

The dataset used provides statistical properties of each video, along with its title and description text. Our very first step was to clean the data and preprocess it. The text was converted to lowercase, tokenized, and then stemmed. We then scraped the thumbnail images of each video to use it later for the analysis. Exploratory Data Analysis was now performed to understand the distributions of features and check for other interesting conclusions that could be drawn from the statistical features. The EDA also involved text features to get an idea for how trending videos were generally titled and described. To understand the thumbnails of these videos, we performed clustering on the colors present in these thumbnails and found the centers around which most colors were present. Finally, we built a deep model that categorized videos into the several available categories using the thumnail, the text, and other statistics of the video.

II. Datasets

The dataset used for analysis can be downloaded here. The data given is for ten countries for the period 2017-2018. Each entry contains 16 attributes describing the video some of which are the title, description, publish time, etc. The images had to be scraped using the thumbnail links provided in the dataset.One of the sample entry is shown below

video_id	date (trending)	title	channel
2kyS6SvSYSE	17.14.11	WE WANT TO TALK ABOUT OUR MARRIAGE	CaseyNeistat
category_id	publish_time	tags	views
22	2017-11-13T17:13:01.000Z	SHANtell martin	748374

...

The other entries are likes, dislikes, comments count, etc. The category corresponding to the category_id was given in a separate json file and was different for each country. For preprocessing the data the following steps were followed.

1) Scraped all thumbnails and downloaded them into a separate folder.
2) Tokenized and stemmed the text fields such as title and description for further analysis and use in prediction model.

Note that the the dataset listed the videos trending on each day and hence there may be multiple instances of the same video as a video may be trending for more than one day. For our analysis we used the data for US, Great Britain and India and performed our analysis on all three countries. However, we have provided the insights mainly about the observations we found for the Indian data. The same can be found in the python notebooks used for the EDA.

III. Analysis Pipeline

A. Exploratory Data Analysis

We begin our analysis with EDA, and look into the characteristics or distribution of each attribute of the data. Several graphing techniques were used to find key insights. We began by simply describing the dataset and finding statistical measures of the numerical columns. Following this, we analyzed the text fields - description and title, by using word clouds and finding the histograms of their length. Several other techniques such as qq plots, correlation matrices, bar plots and histograms were used to draw insights and reach to conclusions that are presented in the Results section IV.

B. Predictive Analysis

For the predictive part of this project, we aimed to predict the category of a video using the data available to us. To be able to use the textual data, we encoded it using the tf-idf vectorizer method whose explanation can be found here [1] . We used convolutional layers for the image part of the data and linear layers for the numerical data(including text which is now in numerical format). Further details about the model can be found in the Results section IV-C.

IV. Results

A. Exploratory Data Analysis

- Most Viewed Videos :
  The most viewed videos for India were :

(a) YouTube Rewind: The Shape of 2017

(b) Marvel Studios' Avengers: Infinity War Official Trailer

Fig. 1: Most viewed videos in India

- The histograms of views and likes of videos published in india is as follows

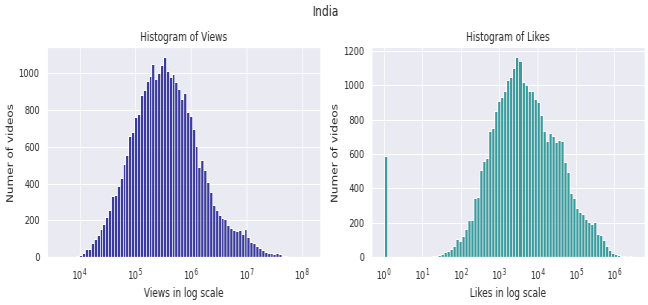


Fig. 2: Histogram of likes and views (India) in log scale

Please refer to the python notebook for the qq plots of these distributions and the histograms for US and Britain. When compared to those of Britain and US we observe that the mean number of views on trending videos is relatively lower in India. The reason for this could be that the videos published in US and Britain have more global viewers, unlike India where the content is more local.

- Hour of Publishing :
  The histogram has two peaks, and when converted to Indian Standard Time, these peaks are at 12pm and 7pm.

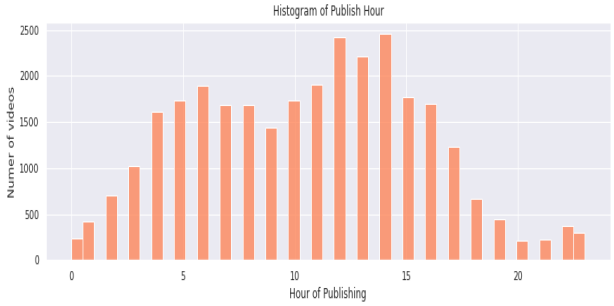


Fig. 3: Histogram of Publish Hours

- Correlation matrix of likes, views, dislikes and comment count. As one would expect all four attributes are highly correlated.

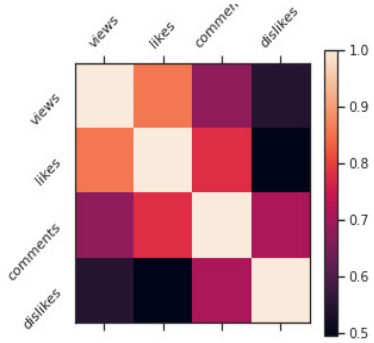


Fig. 4: Correlation matrix

- Word Clouds:
  Word Clouds give an unique insight about the most trending videos and the trending searches for a region. The image below shows the word cloud of the title and description of trending videos in India.

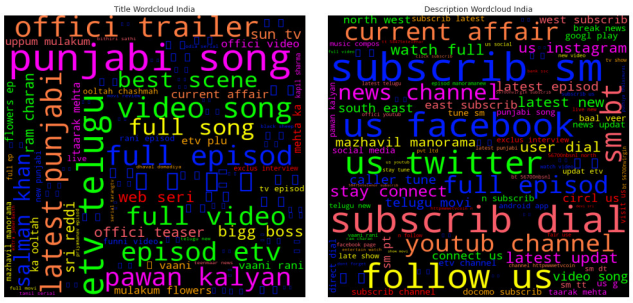


Fig. 5: Title and Description Cloud

- Number of days for which videos trend
  The following figure shows the histogram of the number of days for which a video was trending on YouTube. We can see that the maximum number of days was 16 in case of India.

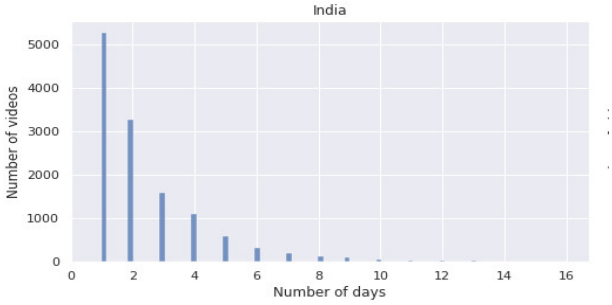


Fig. 6: Title and Description Cloud

- Most Trending Categories: In India, Entertainment is the most trending category followed by News & Politics and Music.

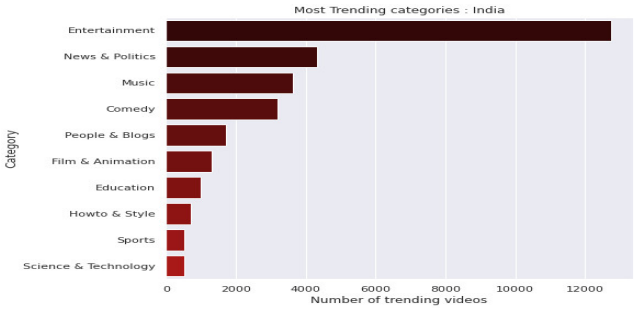


Fig. 7: Title and Description Cloud

- Trendiest Channels: These are the channels that have the maximum number of videos featuring in the trending list.

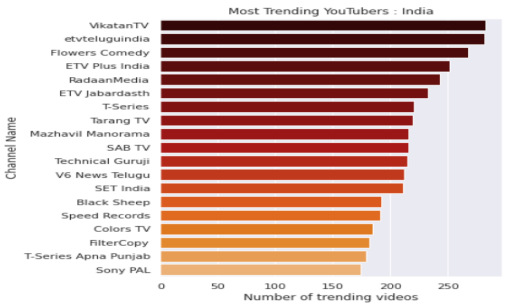
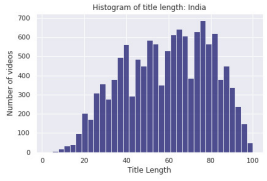
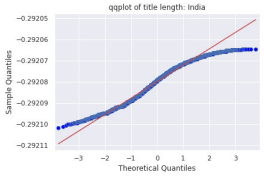


Fig. 8: Title and Description Cloud

- Title Length:



(a) Histogram of title length

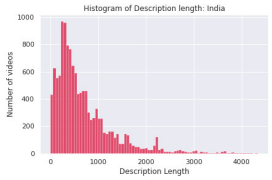


(b) qqplot for title length

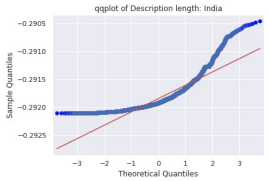
Fig. 9: Analysis of title length

The title length has a distribution similar to a normal distribution with mean close to 60. This is verified by the qqplot adjacent to it.

- Description Length:



(a) Histogram of description length



(b) qqplot for description length

Fig. 10: Analysis of description length

Unlike the title length, the description length has a distribution similar to a gamma distribution.

- Most Viewed Categories: Note that previously we found out the most trending categories. The most trending categories were then decided by the total number of videos in each category in the data frame, whereas the most viewed categories are decided by the categories that attract the maximum number of views. An important observation here is that

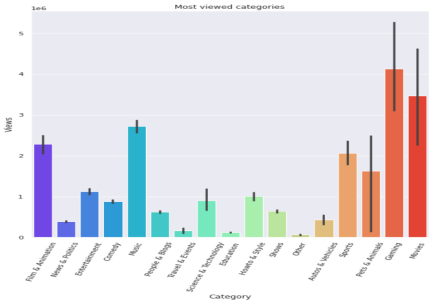


Fig. 11: Most Viewed Categories

although gaming doesn't feature in one of the top trending categories it attracts the maximum number of views.

- Likes to dislikes ratio: Categories such as education have a lot more likes than dislikes, whereas categories such as news and politics have a relatively higher number of dislikes.

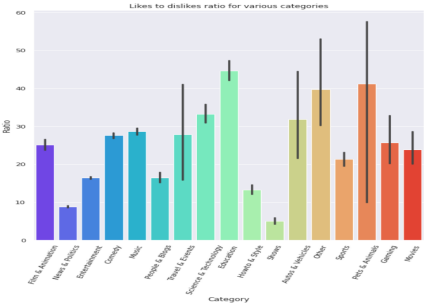


Fig. 12: Likes to dislikes ratio

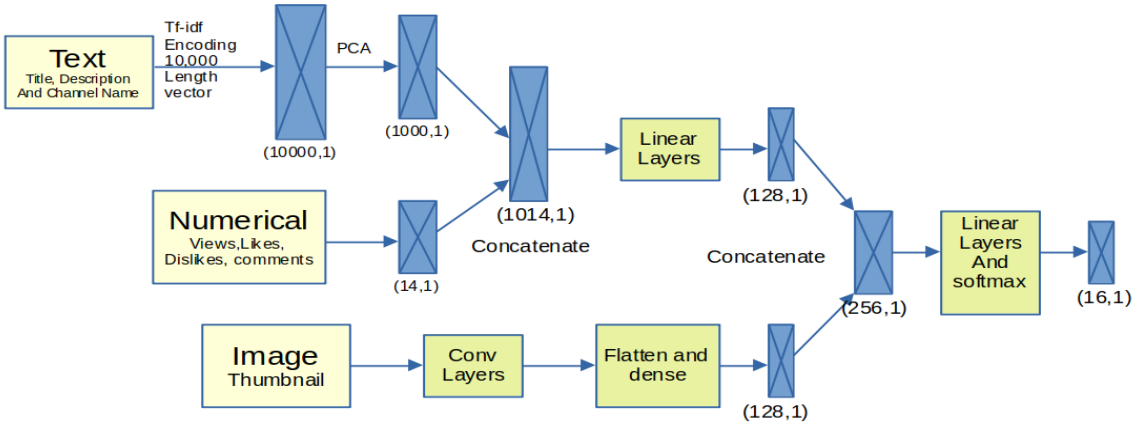


Fig. 13: Description of the model employed

B. Clustering

A batchwise K-Means clustering algorithm was used to find the clusters in the pixel colors of thumbnail images. The centers of these clusters were then plotted in a 3d plot :

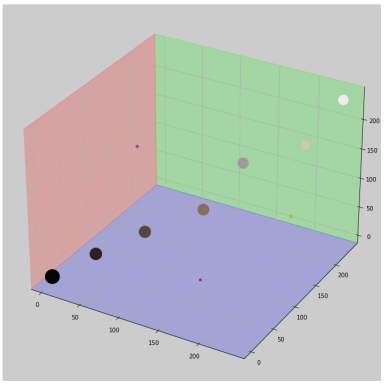


Fig. 14: Kmeans on the pixel colors

C. Predictive Model

The model built is used to predict what category a particular video belongs to by using its text attributes, its thumbnail and other numerical features like views. The data used was of US and Britain. We dropped the duplicates and concatenated the two data frames. Text was preprocessed as has been previously described in 2. This was followed by Tf-Idf vectorization of the text. Unigrams and bigrams were both considered when building the vocabulary for the vectorizer. Post vectorization, the dimension of the text data was reduced (from 10000 to 1000) via PCA, retaining 70% of the variance. This data was concatenated with statistical parameters and then passed through a set of linear layers with ReLU activation. Thumbnails were passed through multiple convolution layers, flattened, and then passed through linear layers with ReLU activation. Both of these were then concatenated and passed through linear layers followed by a softmax layer for classification. Figure 13 is an accurate representation of the architecture used.

The above architecture, with apt dropout between layers, results in test accuracies of over 75%. Other architectures involving just text and statistical features were also tried and the best test accuracy seen was 72%.

While 75-76% may not seem to be very impressive, the explanation for this is that the number of categories were 16, which is high for a classification task and the data available was close to only 8000 entries. Of the 16 categories many are similar, for example movies and entertainment are two very similar categories and the model may, at times, not be able to distinguish between such similar cases. Furthermore, the distributions of the classes is very skewed as has been seen in the EDA part. Yet, the model has F1 scores of over 0.75.

SUMMARY

In this analysis of trending YouTube videos, we have gained insights about their various features and have developed a model capable of categorizing videos without actually watching videos to a great extent.

We evidently see that 'catchy' titles and contrasting colors in thumbnails are frequently present in trending videos. We also looked at the correlation of features and various aspects of specific features too. We further analysed what categories tended to make it to the trending list more frequently and which ones got the most views.

The model developed was able to classify videos well, although it was far from perfect classification. The model was tested and evaluated on metrics such as accuracy and F1 score. We looked into reasons why the model did not give very high scores and tried to help the model generalize(not too rapidly) by using the dropout technique.

We have used various plots like word clouds and histograms, machine learning algorithms like K-Means clustering and Principal Component Analysis and deep learning models involving multiple types of inputs and layers such as convolution, fully connected and dropout.

ACKNOWLEDGMENT

We would like to express our sincere gratitude to the following people:

- Prof. Manjesh Hanawal, Prof. Amit Sethi, Prof. Sunita Sarawagi and Prof. S. Sudarshan for teaching us everything we needed for this project and beyond, with utmost focus on our learning.
- Our Teaching Assistant, Parikshit Bansal for guiding us through the practical aspects of this course.
- Vibhav Aggarwal for assisting us with web scraping techniques and helping us to speed them up.
- Eeshaan Jain and Anupam Nayak for suggesting architectures and methods to improve the models performance.
- Our coursemates for helping us with our doubts and supporting us when we failed to understand concepts.

REFERENCES

[1] TF-IDF vectorizer scikit-learn
https://medium.com/@cmukesh8688/tf-idf-vectorizer-scikit-learn-dbc0244a911a

[2] Multi-Input models transfer learning for image and word tag recognition
https://towardsdatascience.com/deep-multi-input-models-transfer-learning-for-image-and-word-tag-recognition-7ae0462253dc

[3] Training an image classifier (PyTorch)
https://pytorch.org/tutorials/beginner/blitz/cifar10_tutorial.html

[4] Scikit-learn Mini Batch Kmeans documentation
https://scikit-learn.org/stable/modules/generated/sklearn.cluster.MiniBatchKMeans.html

[5] EDA analysis done on the YouTube trending dataset
https://ammar-alyousfi.com/2020/youtube-trending-videos-analysis-2019-us?src=kgl#data-source

[6] Exploring YouTube trending statistics EDA
https://www.kaggle.com/donyoe/exploring-youtube-trending-statistics-eda

[7] Idea about NLP for classification
https://towardsdatascience.com/analyzing-text-classification-techniques-on-youtube-data-7af578449f58