Youtube Trending Videos Analysis

Data Mining Project

https://github.com/MRDVGroupProjectCSPB4502

Vidya Giri	Rouzbeh Salehi Pour	Michael Taylor	Daniel Shackelford
Post-Bacc CS	Post-Bacc CS	Post-Bacc CS	Post-Bacc CS
CU Boulder	CU Boulder	CU Boulder	CU Boulder
Houston, Texas	Vancouver, BC	Boulder, Colorado	Boulder, Colorado
United States	Canada	United States	United States
vigi7924@colorado.edu	rosa7328@colorado.edu	mita6978@colorado.edu	dash6877@colorado.edu

1. PROBLEM STATEMENT/MOTIVATION

Youtube is an American video streaming platform that serves as a key source of news, entertainment, and reference in the digital age: allowing all ranges of global users to broadcast their videos to the world and view videos that are relevant to their interests. *Variety* magazine states that in order, "to compile its trending-video rankings, YouTube uses a proprietary algorithm that factors in total views, likes, comments and searches."[1] This project intends to identify what determines the performance of a trending YouTube video through the analysis of string, numeric, and visual data.

In our project, we have divided the broad topic of characterizing trending videos into four main groups. Firstly, we will analyze whether various tags have an effect on viewership and lead to higher view counts. Secondly, we would like to evaluate whether videos that become abnormally popular in a particular category lead to an overall increase in viewership in the same category. Third of all, we will analyze features that attribute to viewership count such as the video duration, like-to-dislike ratio, video title length, video description length, and date/time posted. Fourthly, we will analyze trending video thumbnails with visual analysis tools to locate clusters of similar thumbnails and prevalence of faces in the thumbnail images.

[1] Spangler, T. (2020, December 1). *YouTube Reveals 2020 Top Trending 2020 Videos*. Variety. https://variety.com/2020/digital/news/youtube-top-trending-dave-chappelle-1234842592/.

2. LITERATURE SURVEY

There are ten documented projects on Kaggle that use the YouTube trending video dataset. Many of them are redundant exploratory data analysis projects. The three projects that appear to be the most relevant to our project is the "Sentiment Analysis USA (Step by step| Emojis|Tags)" project, "Youtube US Trending Videos EDA" project and the "Word Cloud of South Korea YouTube" project.

2.1 Sentiment Analysis USA (Step by step| Emojis|Tags) Project

The project first tried to identify if there was a relationship between view count, likes and dislikes. A linear relationship between the three values was identified. In both scenarios likes and dislikes lead to increased video view counts, but videos performed better if they were getting likes rather than dislikes. The project also identified which video tags were the best performers and displayed the words in a word cloud. Lastly, the project performed emoji analysis, but it is not clear what the creator of the project was trying to accomplish with the emoji analysis.

2.2 Youtube US Trending Videos EDA Project

This project attempts to identify what attributes of videos have a stronger correlation with becoming a high-performing trending video. The project author examined likes, dislikes, comment count, genre, etc., along with if the popularity of the video channel the video was published on had an impact.

2.3 Word Cloud of South Korea YouTube Project

This project is mostly written in Korean text which makes it difficult to understand precisely what the creator of the project is trying to achieve. Based on the source code, it appears that the project creator is using words that are nouns gathered from the video titles, channel titles, video tags or video descriptions to compile a list of words to be displayed in a word cloud.

3. PROPOSED WORK

Because the data set is constantly appended to, the first step of work with the data is establishing a baseline from which to work from. This is to ensure that integration of the solution across all team members is successful. Prior to the final release of the solution, an attempt to pull in all data received during analysis efforts can be made to have the most up to date solution.

Once the dataset is defined, a general skim of the data must be completed to ensure data accuracy and consistency across all of the data. Redundant and duplicate data values will need to be scrubbed, along with any rows that do not contain the relevant data. It is already known that there are some rows that contain null values. In scenarios where the null values can be filled in with placeholders, such as in the video description category, or where insertion of different values, such as the central tendency of the data set, those values will be filled in to preserve as much pertinent information as possible in our data set. In situations where filling in those values does not make sense, the data point may be removed. Data will also need to be scrubbed due to the language barrier. The team will unfortunately not have the requisite language skills to sort through data in other languages, and translation of that data would introduce too many unknowns into the set. All data presented not in english will be removed. Finally, outliers to the set will need to be removed based on standard deviation and overall trend analysis.

In analysis, the data will need to be normalized so that clustering techniques can be used in a number of the tests. In investigating sentiment analysis as it pertains to punctuation, the pertinent information in the dataset will need to be split and parsed so that counts of different punctuation can be captured. In category and tag analysis as it relates to trending duration, data will need to be separated into specific categories, and separate metrics will need to be created for each group. Trending duration denotes time analysis and each category will need to be analyzed over the entirety of its data timeline.

4. DATA SET

The dataset is YouTube Trending Video Dataset (updated daily) from Rishav Sharma on Kaggle.[2] The dataset includes the daily trending YouTube videos from 11 countries with up to 200 listed videos from each day. Some attributes that are included are the video title, channel title, date published, video tags, views, likes, dislikes, description, thumbnail, and comment count. The dataset size is 610,322 rows with 17 attributes.

[2] Sharma, R. (2020). *YouTube Trending Video Dataset (updated daily)*. Kaggle.

5. EVALUATION METHODS

Our evaluation methods will be dependent on the models and methods used to obtain the results. In the case of evaluating supervised learning algorithms, validation curves can assess if a model is overfitting at a specific parameter value. Confusion matrices can provide a detailed look at how well a model predicts each individual input class. For binary classification models, the values from a confusion matrix can be used to construct an ROC curve to measure the separability of the results. The overall performance of a model can be summarized by its F1 score, which is the harmonic mean of its recall and precision scores, allowing for direct comparisons between different models.

When evaluating unsupervised learning algorithms such as clustering algorithms, the elbow method can identify the optimal number of clusters to generate. Furthermore, silhouette analysis can be performed to measure the separability between the clusters. Since the questions we are asking all refer to data that exists in the dataset (i.e. is labeled), all data points within each cluster can be directly compared to their respective labels. This means that the percentage of each label contained per cluster can be calculated to assess the performance of the clustering algorithm and potentially find new patterns within the data.

In all cases, visualizations will be used in both the initial assessment of the raw data and the final results of the models. These can include histograms to visually assess the distribution of the data, correlation heatmaps and pair plots to compare features between each other, scatterplots to visualize clusters (use PCA or t-SNE for models with more than two features), etc.

6. TOOLS

The bulk of work for this project will be done in Python3. During the data cleaning and preprocessing, tools like Pandas and Numpy will be used heavily to create efficient and accurate data frames to feed into subsequent data analysis techniques. Analysis of our data will use scikit-learn (sklearn) as its main additional library, along with the aforementioned tools. This tool was chosen because of the team's prior work experience, and overall ease of use and abundant documentation of the library.

To scrape thumbnail images via url and Youtube video IDs, we will utilize urllib.request to save and analyze images. For face detection analysis, we will utilize OpenCV Haar Cascade classifiers, which is a pre-trained machine learning object detection model for face recognition. We will also use scikit-image and Pillow's Image module for further image handling and processing. For thumbnail clustering analysis, we will utilize the VGG16 convolutional neural network model via tensorflow/keras to extract features. Then we will cluster the top trending videos using KMeans, PCA, and/or t-SNE from sklearn.

Seaborn and MatplotLib will be the main tools used for the creation of charts and data visualizations throughout the project. MatplotLib and Seaborn have been chosen for this project because of their seamless integration with Python, ease of use, and team familiarity.

7. MILESTONES

Project Part	Estimated Completion
2: Proposal Paper	07/04/2021
3: Progress Report	07/30/2021
4: Project Final Report	08/08/2021
5: Project Code and Descriptions	08/08/2021
6: Project Presentation	08/08/2021

8. MILESTONES COMPLETED

The first completed milestone was the Exploratory Data Analysis. This provided valuable insights into the data set that has allowed the team to augment our initial questions with additional constraints, generate additional features, and allowed for the creation of alternative questions in the data set. The next four completed milestones were for each of our 4 main questions.

In answering the first question of tags and viewership, our initial analysis showed a low correlation between tags and total view count of said video. There were no tags of significance as we believe that these are primarily used for search engine optimization and not ranking on the trending page.

Delving into the second question posed in this report, an analysis of the video category as it relates to view count was performed to see if there were any signs of a "ripple effect" in a category after a video in a particular category became more popular.

The third question was answered by assessing the importance of each relevant feature with regards to the view counts. This was completed by placing view counts into a ranking system, predicting each video's view rank using specific features as input data, then assessing the importance score for each feature.

To answer the fourth question, we performed analysis on trending US video thumbnails with visual analysis tools to locate clusters of similar thumbnails in the top trending videos and detected prevalence of faces in the thumbnail images.

9. MILESTONES TO DO

The milestones that are still incomplete are for the final project presentation, project code and descriptions. The code for the project is mostly complete but needs to be refined. For example, some of the words are not written using English text which results in display errors in graphs where the text becomes unreadable. It is also possible that some performance improvements with code run-time need to be implemented.

Analysis of the "ripple effect" phenomenon as it relates to categories and their average viewership numbers will be expanded to the rest of the data set in other countries, and an analysis of the average viewership as it relates to particular categories will be performed to see if the "ripple effect" is only exhibited in particular categories.

Analysis of which features attribute most to view counts needs to be further extended to include categorical and ordinal data. Such features will include the country of origin of the video, day of the week the video was posted, category ID, time of publishing, etc. As long as the supervised learning algorithm can produce a high f1-score on the test dataset, the feature importance scores and their ranking amongst each other can be considered as reliable.

Further analysis on the thumbnails could be done with expanding on analysis and presentation of the detected clusters. It would be interesting to perform further analysis on these new "cluster categories" to see what are the prevalent tags and channels in each cluster and whether these match within certain categories. It would also be interesting to create an information visualization to depict the results of the image analysis in a comprehensive visualization summary.

10. RESULTS SO FAR

Our first task was to identify which tags were correlated with higher video view totals. The mean, median and frequency of video views related to each tag's use were identified. It

appears that tags related to music videos were correlated with high video view counts based on the trending YouTube data we were provided with. However, there does not appear to be a coherent set of rules that would determine effective tags to use for videos to increase video views. This would be consistent with existing research that states that tags have very little impact with video search results and video rankings.[3]

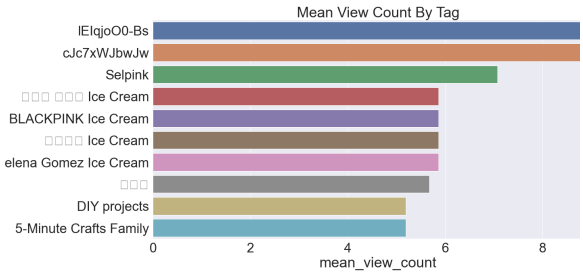

Fig. 1: Correlation between median total video views per video tag.

Delving into the second question posed in this report, an analysis of the video category as it relates to view count was performed to see if there were any signs of a "ripple effect" in a category after a video in a particular category became more popular.

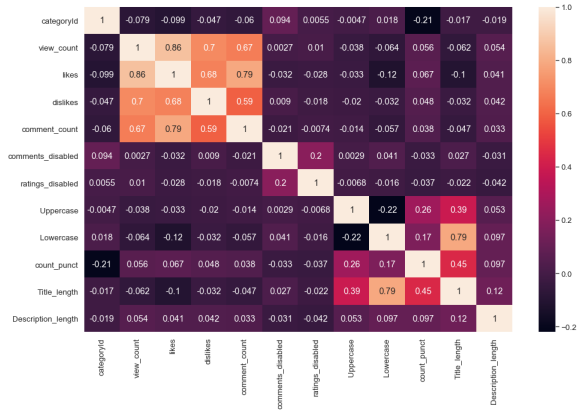

Fig. 2: Correlation matrix of features in the data set as they relate to each other.

To begin this analysis, our data frame was first sorted into values according to the publish date of the video. The data frame was then pruned to display only videos in a particular category. For the initial investigation, the category of "10" was chosen as it is one of the most popular categories in the entire data set. The z-score for the given category was then computed. The z-score was used to determine the number of standard deviations below or above the population mean a raw score is. Z-score is typically used to remove outliers of a given data set, however in this case, the z-score is used to determine which data points had an abnormally large view count. All videos with a z-score higher than 3, or 3 standard deviations from the mean, were indexed and set aside as possible locations of data that may impart a "ripple effect" on videos of the same category.

For each of these "spike points" in the data set, videos of the same category were found that were published 7 days before and 7 days after the time of the video. It is reasonable to assume that a legitimate increase in the viewership of a particular category would see a jump in the average viewership of that category after the video's publishing date. The averages of the videos published 7 days prior to the publishing date of the video and 7 days after the video publishing date were then compared. For the data in category "10" of our video set, it was found that 99 videos show an increase in average viewership and 88 did not. This is an 12.5% increase overall.

The same steps were taken over each of the 15 categories in the data set. After running this analysis, the following results were obtained:

[3] *What are YouTube Tags (and Do They Even Matter)?* https://www.webfx.com/blog/social-media/what-are-youtube-tags/

Table 1: Category ID, 7 day Increased Average, and 7 day Decreased Average for each "spike" in data for the given category

Category	Increased Average	Decreased Average
1	33	17
2	7	9
10	99	88
15	4	6
17	57	47
19	0	6
20	80	73
22	29	34
23	36	26
24	126	145
25	29	10
26	9	5
27	11	16
28	33	26
29	0	0

An observation of the results shows that a total of 8 categories had an overall increase in viewership and 7 categories that did not. Moving forward, data from across all countries will be analyzed for additional clarity on after effects of this phenomenon, and averages across each category will be analyzed to see if this "ripple effect" is only exhibited in some categories.

In analyzing our third question, it was decided a more in depth correlation check would be required. Although standard Pearson correlation coefficients between different features and view count can be calculated, standard Pearson correlation coefficient values do not consider the correlation between the different features. It was decided that to identify the features which best attributed to view count, we would build a machine learning model to classify view counts and assess the model's feature importance.

A random forest model was used in our analysis due to its strong use in Kaggle competitions and possibility to visualize one of its decision trees. In order to use random forest classified categorical data, the view counts feature needed to be mapped to a ranking system based on count of videos per rank and popular milestones (i.e. 500k, 1 million, 10 million, etc.). Select numerical features were selected as the input and a random forest model was created. Cross-validating the dataset through 5 folds, the average f1-score obtained was 90%.

The importance score for each feature is plotted in Figure 3. The importance score is based on the Gini criterion, where essentially the more "impure" or mixed a set of labeled data points are after a split by the introduction of a feature, the lower the importance score for that feature is. As observed, the likes, comment count, and dislikes attribute most to the prediction of a video's view count ranking. In our EDA, we noticed that many trending videos use up to 100 tags for their videos. However, from the data below we can conclude that it is one of the least important features for view count rank prediction, suggesting that more tags on a video does not equate to higher view count when considering other features.

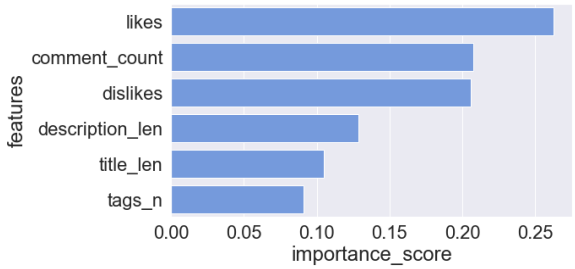

Fig. 3: Random forest model feature importance.

For the fourth question, we performed analysis on trending US video thumbnails with visual analysis tools to locate clusters of similar thumbnails in the top trending videos and detected prevalence of faces in the thumbnail images. To scrape thumbnail images via url and Youtube video IDs, urllib.request was utilized to save and analyze images. Then, these images were used for face detection analysis and clustering analysis.

For the face detection analysis, OpenCV Haar Cascade classifiers were utilized on the thumbnails, returning if a face was detected in the thumbnail or not. This information was then appended to the original data frame which could be utilized for analysis with other variables. From our analysis we have determined that 63.8% of the trending videos in the US contained a detectable face in the thumbnail.

Fig. 4: Youtube video thumbnail depicting face detection analysis with a yellow box drawn demarking the detected face using the OpenCV frontal face Haar Cascade classifier.

For thumbnail clustering analysis, the VGG16 model was utilized through tensorflow/keras to extract features for the top 1000 videos. Once we removed the videos with no accessible public video, this left 945 thumbnails. The components were then reduced from 4096 classifier vectors to 100 using PCA (principal component analysis). Then a plot of the square distance and the number of clusters was utilized to determine the optimal number of clusters with KMeans. By locating the elbow on the plot, 20 was selected as the number of clusters. Then, by using KMeans, 20 groups were classified with the breakdown shown in Figure 5.

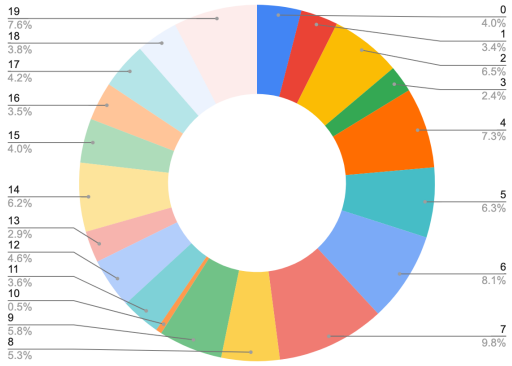

Fig. 5: Pie chart showing breakdown of percent of thumbnails in each of the 20 clusters.

Inspecting some of the clusters such as cluster 18 (Figure 6), we can see that similar trends arise where in this case there is a prevalence of colorful cartoon and animated features in the thumbnails.

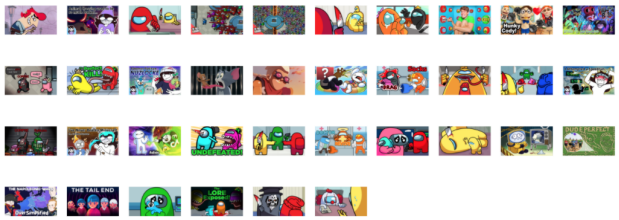

Fig. 6: Summary of Youtube thumbnail images in cluster 18.

Finally, a t-SNE plot (Figure 7) was generated with the top trending US video thumbnails using the detected feature vectors which clustered similar images together, depicting the classification of many images in a visual manner.

Fig. 7: t-SNE plot of Youtube thumbnails with a magnified view of a cluster of videos with similar faces in the thumbnail.