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. ABSTRACT

The overall goal of this project was to analyse youtube trending video data in order to determine data correlations related to an increase in views of particular videos. There are 4 main interesting questions related to this goal. Firstly, do various tags have an effect on viewership? Secondly, do popular trending videos in a particular category lead to an overall increase in viewership in the same category? Thirdly, what features cause an increase in viewership count for a trending video? Fourthly, what are common visual features in trending video thumbnails.

After analyzing these four questions we found that tags have a weak correlation or are not correlated at all to video view count totals. We also found that there was not a direct correlation between an increase in views in a category after a popular video was released, however different categories do experience more of a change in views than others. The feature importance scores from the random forest model showed that likes, comment count, and dislikes were the most important features when predicting video view count and therefore could be the most contributing factors for an increase in viewership. We were able to detect significant visual features in Youtube thumbnails with facial and object detection models and subsequently created 20 defined clusters.

2. INTRODUCTION

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 questions. These questions are significant since they allow us to identify factors and features that determine the performance of a trending YouTube video. The four questions are summarized below:

1. Do various tags have an effect on viewership?

[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/.

We have analyzed whether various tags have an effect on viewership and lead to higher view counts.

2. Do popular trending videos in a particular category lead to an overall increase in viewership in the same category?

We have evaluated whether videos that become abnormally popular in a particular category lead to an overall increase in viewership in the same category.

3. What features cause an increase in viewership count for a trending video?

We have analyzed features that attribute to viewership count such as the number of likes and dislikes, comment count, video title length, video description length, etc.

4. What are common visual features in trending video thumbnails?

We have analyzed trending video thumbnails with visual analysis tools to locate clusters of similar thumbnails and prevalence of faces in the thumbnail images.

3. RELATED WORK

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.

3.1 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.

3.2 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.

3.3 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.

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,

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

description, thumbnail, and comment count. The dataset size is 610,322 rows with 17 attributes.

5. MAIN TECHNIQUES APPLIED

5.1 General Data Cleaning, Processing, Warehousing, and Visualization

The bulk of work on this project was done with Python3 via various Jupyter notebooks in order to annotate and organize analysis that was being done.

During the data cleaning and preprocessing, tools like Pandas and Numpy were used to create efficient and accurate data frames from the source csv files from various regions to feed into subsequent data analysis techniques. Analysis of our data used scikit-learn (sklearn) as the main additional library, along with the aforementioned tools.

To scrape thumbnail images via url and Youtube video IDs, we have utilized urllib.request to save and analyze images. Since some of the videos were now not publicly available via video ID, this was taken into consideration and such data points were dropped for the visual analysis. Finally, Seaborn and MatplotLib were the main tools used for the creation of charts and data visualizations throughout the project.

5.2 Tag Analysis Techniques

In order to address the first question surrounding video tags, a Pandas dataframe first had to be created. The first dataframe only contained video tags and video view count totals. The data frame was then converted to a list of Python tuples. After the list of tuples was created, a giant Python dictionary was created with the keys in the dictionary being the tag name and the value being a list of video view count totals for each video where the tag was used. The tags for each video first had to be split into a smaller list. Once the giant dictionary containing the tags was created a new data dictionary was created to create a new dataframe that contained the list of tags, frequency of tag use, total video view count for each time that each tag was used, mean view count for when each tag was used and median video view count for when each tag was used. The data was then used to create four bar plots using the Seaborn Python library and a word cloud was created displaying tags that were the most frequently used.

5.3 Category Popularity Techniques

The analysis for the second question hinges on one main calculation. This calculation is the z score of the data. In analyzing the data set by z-score, we are directly comparing the current data point to the mean of the data and determining how many standard deviations the data point is from this mean. This allows us to classify what data points are largely different from the rest of the data. Normally this type of analysis is used in an eda to weed out outliers of the data, however in the applications in this paper, and due to the relatively clean nature of the data, the z-score allows us to analyze points of data that remain in the set, but may have a larger effect on the other data points in the set than mean data points.

5.4 Viewership Feature Analysis Techniques

The third question could not simply be answered by analysing the Pearson correlation coefficients between view counts (target feature) and other features (input features) because Pearson correlation coefficients only state the linear correlation between two features and not the correlation between input features. Not considering the relationship between input features may result in incorrectly identifying an input feature as highly contributing to view counts where it could simply be highly correlated with another input feature that is also identified as having a high contribution to view counts. To work around this, it was decided that

we would build a machine learning model to classify view counts and assess the model's feature importance scores.

The idea is that if we were to predict video view counts with a high accuracy, then the feature importance scores of the model could provide a good indication of what features contribute most to view counts. A random forest model was used in our analysis due to its strong use in Kaggle competitions and its ability to handle large data. Since a random forest is a classification model, the view counts would have to be placed in bins/ranks. Ranks were chosen based on the count of videos per rank and popular viewership milestones. The count of videos per rank are shown in Table 1.

Table 1: Rankings/bins of videos by minimum view count.

Ranks	Minimum view count	Video count
A	10,000,000	17,011
B	5,000,000	21,500
C	2,500,000	43,681
D	1,000,000	103,746
E	500,000	99,101
F	250,000	91,161
G	100,000	79,625
H	0	44,049

5.5 Thumbnail Visual Analysis Techniques

For question 4, a variety of techniques were used for face detection analysis and image clustering analysis.

For face detection analysis, we utilized OpenCV Haar Cascade classifiers, which is a pre-trained machine learning object detection model for face recognition. Using this technique, we were able to add an additional feature to the data frame to record if a face was detectable, with an array denoting the coordinates of each identified face within the image that was analyzed.

For the image clustering analysis, we used scikit-image and Pillow's Image module for further image handling and processing. For data persistence, Pickle was used. Additionally for the thumbnail clustering analysis feature detection, we have implemented the VGG16 model via tensorflow/keras to extract feature vectors.

VGG16 is a convolutional neural network that was proposed by K. Simonyan and A. Zisserman of the Visual Geometry Group Lab of Oxford University in 2014 in the paper "Very Deep Convolutional Networks for Large-Scale Image Recognition".[3] We selected this model due to its high accuracy in classifying a wide variety of image features and objects. Because of these qualities, the VGG16 model can be applied to many situations such as Youtube thumbnail analysis.

Overall, using this classification model, we could then cluster and organize the top trending videos using KMeans, PCA, and t-SNE from sklearn.

6. KEY RESULTS

6.1 Tag Analysis Results

The first question that needed to be answered was whether various tags have an effect on viewership and if they lead to higher view counts. The analysis was straightforward. The use of tags appeared to be weakly correlated with video performance which is consistent with YouTube video search engine optimization (SEO) research.[4] It is possible that tags have no measurable impact on video view count at all. However, we were able to visualize some of the common tags in a word cloud (Figure 1).

[3] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556.*

[4] Bonacci, J. (2021). *What are YouTube Tags (and Do They Even Matter)?* WebFX Blog. https://www.webfx.com/blog/social-media/what-are-youtube-tags

Fig. 1: Word cloud visualizing of common trending video tags

The mean (Figure 2) and median (Figure 3) view totals for each video tag appeared to be the most relevant when evaluating tag video view performance and no useful information could be observed regarding the use of tags that could be repeatable. It appears that there was a correlation with videos having tags and having high median and mean video view counts, when compared to having no tags at all however. Interestingly enough, the video tags that were the most frequently used did result in the highest mean and median video view count totals.

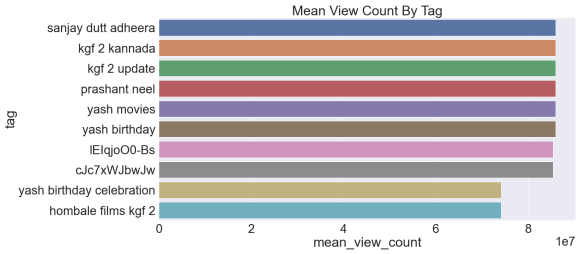

Fig. 2: Mean View Count as it relates to tag

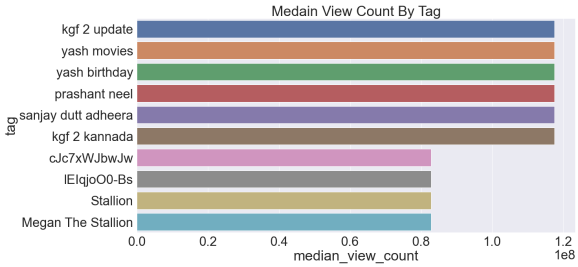

Fig. 3: Median View Count as it relates to tag

6.2 Category Popularity Results

Delving into the second question posed in this report, an analysis of the video category as it relates to view count was performed in order to see if there are any signs of a "ripple effect" in a category after a particularly popular video is published.

To begin this analysis, the data frame was first sorted into values according to the publish date of each video. The data frame was then pruned to display only videos in a particular category. For the initial investigation, the category for music (music has an id of 10 in the data set) was chosen, as it is one of the most popular categories in the entire data set and contains the majority of the top 10 most popular videos on the platform. The z-score for the 'view_count' feature was then calculated for the given category in order to find outliers in the data set. The z-score represents a measure of standard deviations below or above a population mean a specific data point in the 'view_count' feature a video is. Z scores are typically used for outlier analysis in order to determine and throw out outliers in the data set, in this case however, the z-score is used to determine data points that have an abnormally high amount of views. This can be done as the data set used has not only already been cleaned, but is also inherently a very accurate data set. In the end, all videos with z-score higher than 3, or more precisely 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. It is interesting to note that the total analysis was computed with different z-scores (lower z-scores equating to more total data points of interest), however after analysis, similar results were seen. This led to the establishment of a z-score of 3 as the cutoff as this also coincides with common data analytics practices.

For each of the points of interest in the category, or more specifically 'view count spikes', videos published in the same category 7 days prior to the publishing date of the video and 7 days after the publishing date of the video were found. 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 published date. The average of the category videos published up to 7 days prior to the video was calculated, followed by the average of the category videos published up to 7 days after the videos publishing date. For the music category in question, 657 videos saw an increase in average views from prior to post week, however 702 videos saw a decrease in average views from prior to post week. This equates to a 6.8% higher chance of a decrease in views.

The same steps were taken over the remaining 33 categories in the set. After running this analysis, the following results were obtained:

Table 2: 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	158	127
2	41	43
10	657	702
15	21	30
17	391	617
18	0	0
19	62	7
20	349	411
21	0	0
22	350	264
23	228	295
24	1063	989
25	157	137
26	76	30
27	118	116
28	182	170
29	4	0
30	0	0
31	0	0
31	0	0
33	0	0
34	0	0
35	0	0
36	0	0
37	0	0
38	0	0
39	0	0
40	0	0
41	0	0
42	0	0
43	0	0
44	0	0

An observation of the results shows that a total of 9 categories saw a positive increase in viewership, 6 categories saw a negative increase in viewership, and 18 categories did not exhibit a large enough change in viewership. We can plot the top average increases as follows:

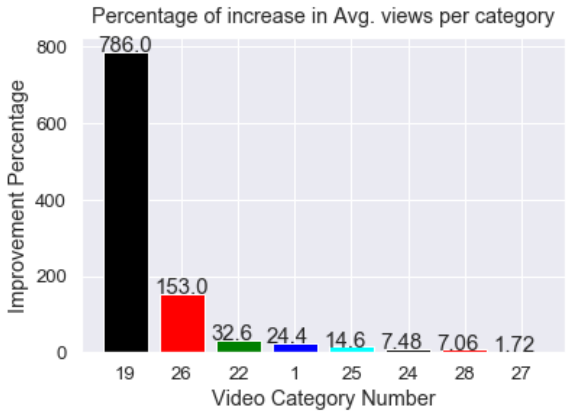

Fig. 4: Percentage increase in in average viewership per category

With the top decreases in average viewership in these categories:

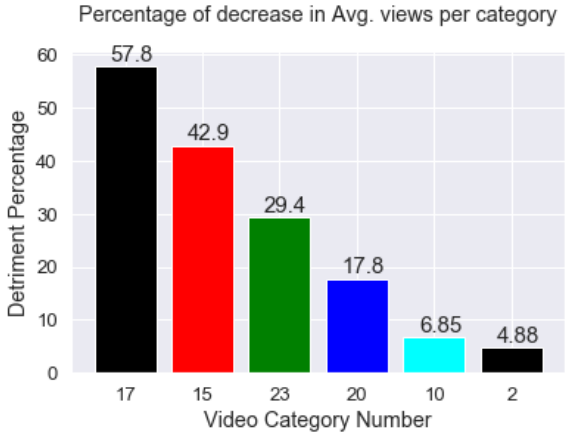

Fig. 5: Percentage decrease in average viewership per category

Although there is not a strong enough correlation in all of the categories to unequivocally denote that the 'ripple effect' is certainly exhibited, it does seem to have a greater effect on certain categories. For instance category 19 denotes videos related to travel and current events. A large increase in viewership after a popular video in this category makes sense as world events and news fit into this category. While category 26 denotes videos in the 'gaming' category. A direct causation would need to be investigated further in this category, but it stands to reason that gaming sees a large increase in views when notable events or products are exhibited/fall into the spotlight. The top three categories that see a decrease in average viewership are Sports, Pets and Animals, and Comedy respectively. Again, further causation would need to be investigated further, however these categories all share a similar stance of personal entertainment.

6.3 Viewership Feature Analysis Results

Before creating the model, additional features were developed within the EDA process to observe if certain tactics used by YouTubers to increase viewership on their videos would actually be effective. These features include the number of tags associated with a video, the title length, number of upper and lower case letters in the title, number of punctuations in the title, etc. The final model had 11 input features, all of which were numerical.

Initially, the model was to be built to include categorical data as well as numerical. The categorical data would include category ID, country, and the month, hour, and day of the week the video was published. However, during the implementation of the model, there were some issues that were encountered. First, a random forest model can only use categorical input features if the feature is one-hot encoded or dummy variables are created for each feature's values. Using this method created 73 additional features for the dataset, making the model more complex than needed. Second, the Wikipedia article on random forests states that for "categorical variables with different number of levels, random forests are biased in favor of those attributes with more levels" when calculating feature importance. Third, if we were to include the categorical features in the model and determine their feature importance, we would be dealing with 84 feature importance scores. Even the categorical feature we were most interested in, category ID, had 15 unique values. Due to these reasons, it was decided to only use numerical data for the model.

With the 11 numerical input features, the data was normalized using a min-max scale between 0 and 1. The model was fitted to a training dataset of ~350,000 data points and predicted the test dataset of ~150,000 data points, yielding an f1 score of ~87%. A cross validation score using 5 folds was also calculated using the random forest model and the entire dataset of ~500,000 data points, providing a mean f1 score of ~88%. A normalized confusion matrix showing the recall score for each class in the output feature is shown in Figure 6. The recall scores were very similar to the f1 scores for each class.

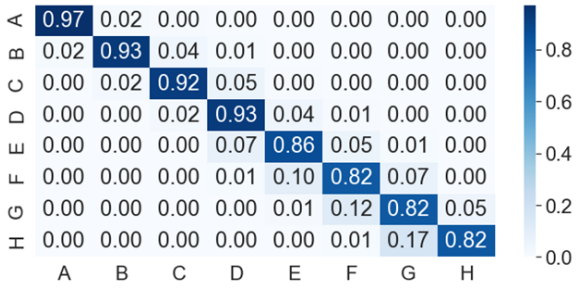

Figure 6: Confusion matrix for the random forest model.

The higher ranked videos (videos with more views) have a better classification score than the lower ranked videos. One reason for this could be that since the lower ranked videos have a larger quantity of videos associated with them,

the variability between these videos may be higher compared to the higher ranked videos that have less videos, making it more difficult to correctly classify them. Nonetheless, an overall f1 score of ~88% is very acceptable, therefore the feature importance scores of the model can be accepted and analysed.

The importance scores of all 11 input features are plotted in Figure 7. 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. The plot shows that likes, comment count, and dislikes are the most important feature for predicting the view count rank. Many youtubers always ask viewers to like their videos and leave a comment, and this may be that YouTube's algorithm shares more videos with higher likes or more discussions. This is consistent with the reason given by many YouTubers when asking the user to "like and subscribe". However, it could be the case that as viewership increases, naturally so does the number of likes, dislikes, and comments increase. As mentioned in class, correlation does not equal causation, however we can agree these features are the most important numerical feature in predicting the view count rank and therefore could be contributing to view count.

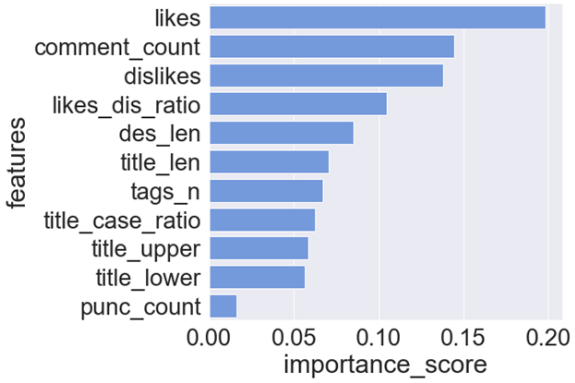

Figure 7: Feature importance scores for random forest model.

Observing that the casing and number of punctuations in the video title were the least important was most surprising. We have all observed YouTube video titles with lots of capital letters and exclamation points in an effort to get the users attention, however it seems that this is not an important element in a video's view count when considering all other input features. The number of tags in a video are also not a significant feature in comparison to the top three mentioned before, which may indicate that users may not be searching for specific topics on YouTube and may be viewing videos that are recommended to them instead. This opinion is further supported by the fact that the number of tags attribute has almost no linear correlation with any other feature.

6.4 Thumbnail Visual Analysis Results

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. This depicts that the majority trending videos do have a visible human face in the thumbnails, and this is a significant feature that may attract users to view a video.

Fig. 8: 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 (Figure 9), 20 was selected as the number of clusters.

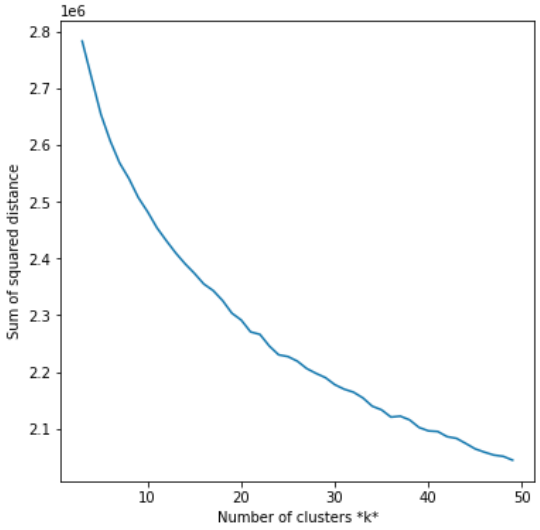

Fig. 9: Plot of the square distance and the number of clusters was utilized to determine the optimal number of clusters.

Then, by using KMeans, 20 groups were classified with the breakdown shown in Figure 10.

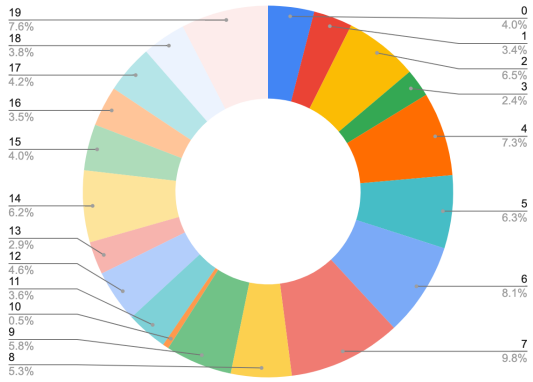

Fig. 10: Pie chart showing breakdown of percent of thumbnails in each of the 20 clusters.

Inspecting some of the clusters such as cluster 18, 13, and 12 (Figure 11-13), we can see that similar trends arise. In cluster 18, we can see prevalence of colorful cartoon and animated features in the thumbnails . In cluster 13, we can see thumbnails with colorful bursts of light that are focused on dramatic music videos and movie trailers. In cluster 12, we can see predominantly sports highlights for soccer and basketball and dynamic human figures.

Overall, by identifying clusters from visual analysis of Youtube thumbnails, we are able to extrapolate trends and “micro-categories” which might exist in larger categories, trends, and creators.

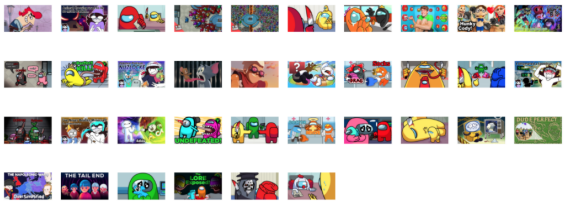

Fig. 11: Summary of Youtube thumbnail images in cluster 18 depicting colorful cartoon and animated features.

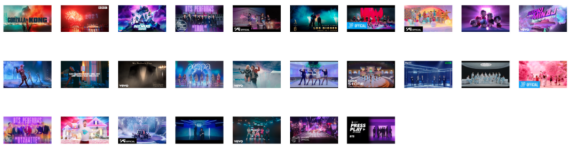

Fig. 12: Summary of Youtube thumbnail images in cluster 13 depicting music videos and movie trailers with colorful bursts of light.

Fig. 13: Summary of Youtube thumbnail images in cluster 12 depicting dynamic human figures and sports highlights.

Finally, a t-SNE plot (Figure 14) 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.

A t-SNE grid (Figure 15) was also created to view similar results in a more linear fashion to depict all thumbnails analyzed.

Fig. 14: t-SNE plot of Youtube thumbnails with a magnified view of a cluster of videos with similar faces in the thumbnail.

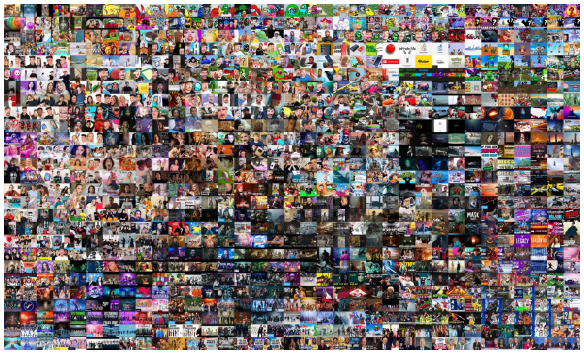

Fig. 15: t-SNE grid of Youtube thumbnails.

In the end, by using feature detection to identify clusters and similarity in Youtube thumbnails, we are able to locate “micro-categories” that are significant within the broader aspect of larger categories, trends, and creators.

This could potentially be used to provide recommendations of similar videos based on the visual features of a thumbnail, or show commonalities among the larger picture. Evidently, since Youtube is used in a variety of contexts, it would be interesting to look into a wider variety of videos beyond the trending set to see if more patterns can be discovered and organized.

7. APPLICATIONS

Overall this analysis is directly invaluable to two main groups, creators wanting to grow their viewer base, and YouTube’s engineers wishing to increase user engagement. Creators can use this information to understand what factors into video success and what they can do to improve their viewers' engagement in their content. YouTube can use this information not only to better their creators, but can also use it to determine server loads and site use, as well as determine which categories of videos are most important to their user base and which categories possibly need more support from the platform.

Looking further out, this data can be used in the design of other social media systems and in

analyzing population response to things like news and its effect on society.

8. VISUALIZATION

To visualize our thumbnail clustering analysis, we have created a viewer that allows the user to interactively explore the visual analysis of the top US trending videos through 3 modes:

1. **Cluster Exploration:** Visualize the micro-categories that have been created from the visual clustering of video thumbnails.
2. **t-SNE Plot:** Interactive plot that visually clusters similar video thumbnails on a 2-D plot with t-SNE methods.
3. **t-SNE Grid:** Interactive grid that visually depicts similar video thumbnails in an organized grid view with t-SNE methods

The visualization can be accessed at: https://mrdvgroupprojectcspb4502.github.io/