
# Team 28: Trending Youtube Video Analysis

Mitali Bante
Tracy Cui
Ashutosh Rawat
Wei-Lun Tsai
David Ma

# Initial data exploration, Cleaning and Sampling

- The complete dataset has video information for USA, Britain, Germany, Canada,France, Russia, Mexico, South Korea, Japan and India
- We plan to majorly **focus on US video set** because a similar analysis can then be replicated to other countries
- For US, we have a total of **40946 rows** and **16 columns**.
- The data has category_id as 1,2,..44. These values are mapped to their corresponding categories using an additional **json file** provided with the data: eg 1 maps to ‘Film & Animation’, 2 maps to ‘Autos & Vehicles’, 10 maps to ‘Music’ and so on
- The missing values can be seen to the right: only column description has 570 missing values out of 40946 rows.
- The date time in the dataset is converted to the correct Datetime format in order to extract date, month, year and time.
- Some videos recur as trending videos from time to time. The total number of unique videos in our dataset are ***6351***.

**Missing value Analysis**

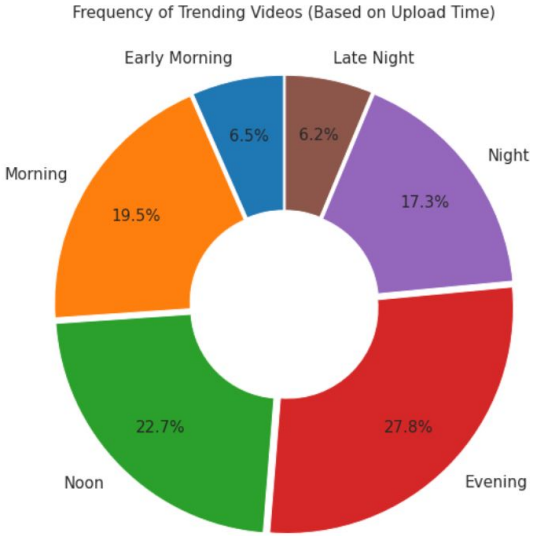
| | |
|---|---|
| video_id | 0 |
| trending_date | 0 |
| title | 0 |
| channel_title | 0 |
| category_id | 0 |
| publish_time | 0 |
| tags | 0 |
| views | 0 |
| likes | 0 |
| dislikes | 0 |
| comment_count | 0 |
| thumbnail_link | 0 |
| comments_disabled | 0 |
| ratings_disabled | 0 |
| video_error_or_removed | 0 |
| description | 570 |

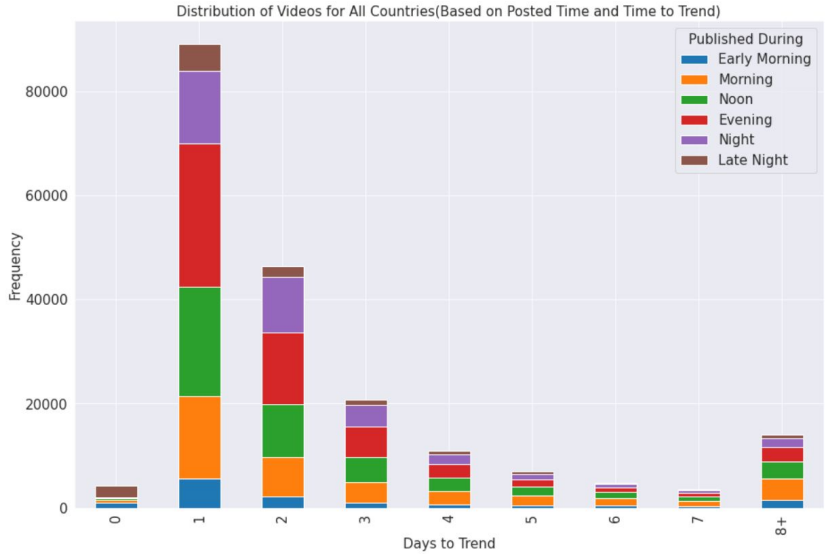
# Feature Analysis

- NLP-related features: Title, Channel_title, Tags, Description
- CV-related feature: Thumbnail_link (Preview image of the video)
  - Use ResNet50 to do object detection
    - No thumbnail (or not available at this moment)
    - No object is detected in the thumbnail
    - Objects are detected in the thumbnail
    - Duplicated thumbnail
- Statistics-related features: Time, Date, Likes, Dislikes, Comment_counts
  - Months (July - Oct) were removed due to inadequate data
  - Emphasis was on most recent years (2017-2018)
  - Time zone conversions to EST

# Data Exploration: Time of watch

- Evening, noon, morning is the best time for videos to be released.These publishing times also show that the trending duration for these videos is long.
- Early morning and late night uploads are rarely watched and have less views compared to other times once trending.



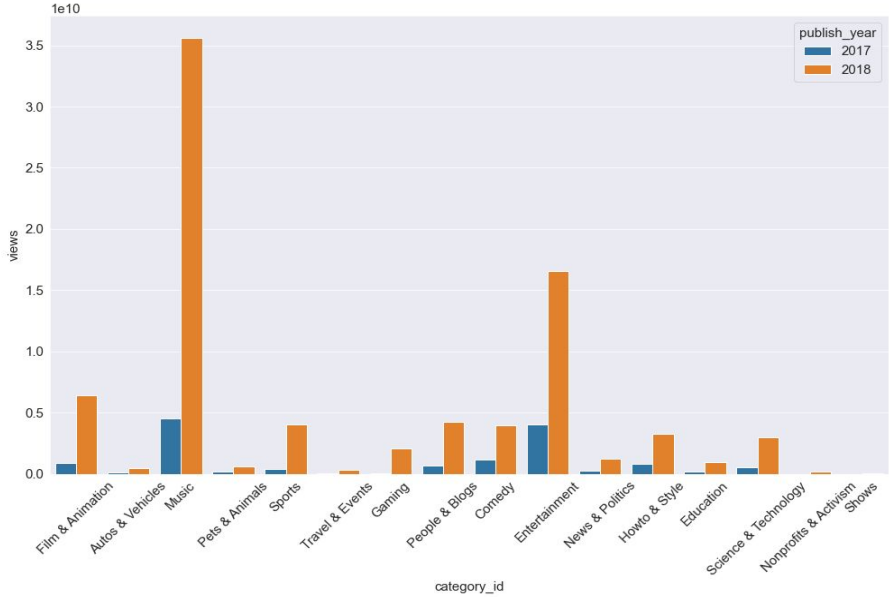
# US specific Data Exploration: Trending video

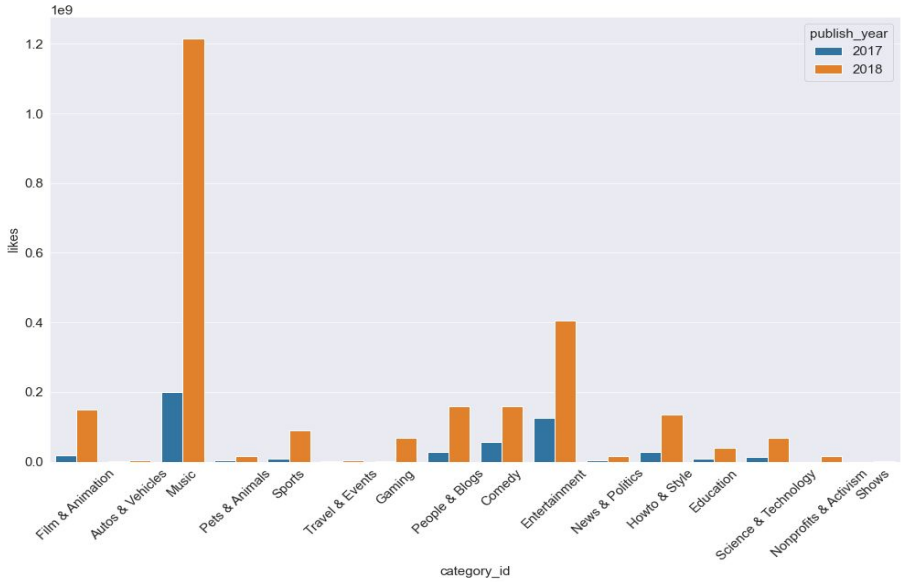
Distribution of views, likes, dislikes and comments with respect to published time


- Most video activity is observed in the Spring (April - June)
- Data under focus is only from Nov 17-Jun 18. Hence the plots only show the corresponding dates
- Likes, dislikes, comment counts all correlate to the views
- There exists as abnormal spike in the number of dislikes in Dec 2017 and Jan 2018
- Views and likes are the most interacted metrics

# US specific Data Exploration: Popular video types

- Music and Entertainment are the most interacted videos (Likes and views)
- Shows and Nonprofits & Activism videos are the least popular
- There is a huge difference from 2017 to 2018 in all categories as the data under consideration is for 2 months of 2017 and 6 months of 2018



# US specific Category distribution of Trending videos

Categories and number of **Videos Published**

| | |
|---|---|
| Entertainment | 9964 |
| Music | 6472 |
| Howto & Style | 4146 |
| Comedy | 3457 |
| People & Blogs | 3210 |
| News & Politics | 2487 |
| Science & Technology | 2401 |
| Film & Animation | 2345 |
| Sports | 2174 |
| Education | 1656 |
| Pets & Animals | 920 |
| Gaming | 817 |
| Travel & Events | 402 |
| Autos & Vehicles | 384 |
| Nonprofits & Activism | 57 |
| Shows | 57 |

Categories and number of **Likes**

| | |
|---|---|
| Music | 1416838584 |
| Entertainment | 530516491 |
| Comedy | 216346746 |
| People & Blogs | 186615999 |
| Film & Animation | 165997476 |
| Howto & Style | 162880075 |
| Sports | 98621211 |
| Science & Technology | 82532638 |
| Gaming | 69038284 |
| Education | 49257772 |
| Pets & Animals | 19370702 |
| News & Politics | 18151033 |
| Nonprofits & Activism | 14815646 |
| Travel & Events | 4836246 |
| Autos & Vehicles | 4245656 |
| Shows | 1082639 |

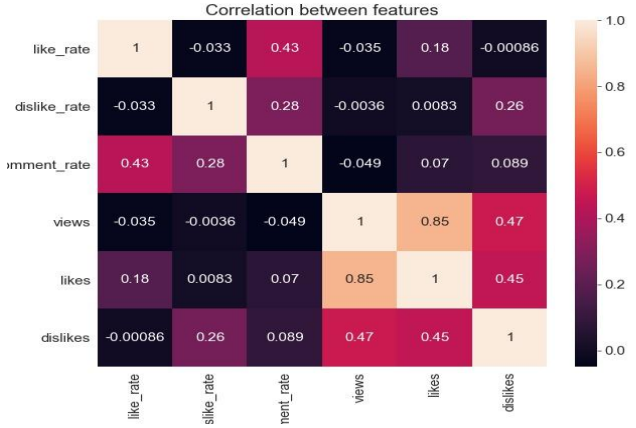
Categories and number of **Views**

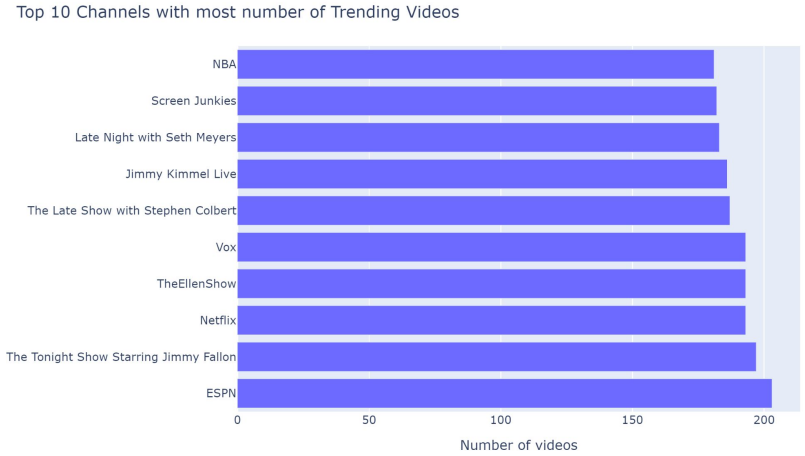
| | |
|---|---|
| Music | 40132892190 |
| Entertainment | 20604388195 |
| Film & Animation | 7284156721 |
| Comedy | 5117426208 |
| People & Blogs | 4917191726 |
| Sports | 4404456673 |
| Howto & Style | 4078545064 |
| Science & Technology | 3487756816 |
| Gaming | 2141218625 |
| News & Politics | 1473765704 |
| Education | 1180629990 |
| Pets & Animals | 764651989 |
| Autos & Vehicles | 520690717 |
| Travel & Events | 343557084 |
| Nonprofits & Activism | 168941392 |
| Shows | 51501058 |

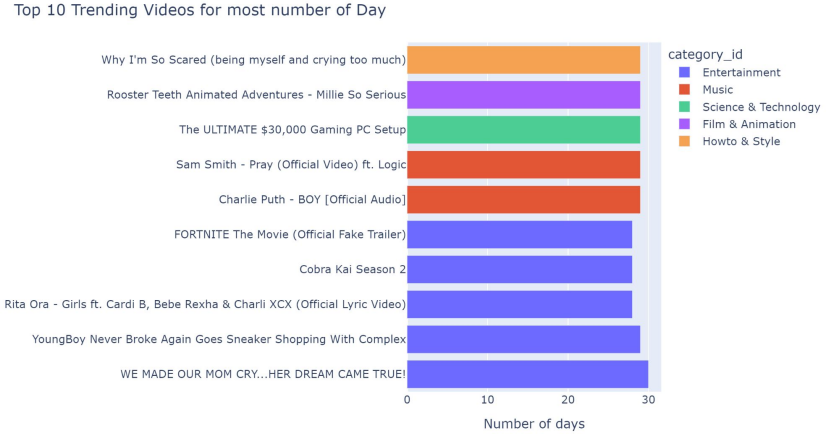
Music and entertainment have the highest number of videos as well as likes and views but ‘How to & Style’ category of videos have lesser number of views and likes when compared to the number of published videos. Reversely, ‘Film & Animation’ have more likes and views compared to the number of videos released

# US Specific Correlation and some Interesting charts

It is known that likes will directly depend on views. This can be seen from the correlation chart too. But the like_rate ((like/views)*100), dislike_rate and comment_rate show no dependance on the number of views. So it can be concluded that popularity (more number of views) and like_rate (like conversion rate) is not correlated.



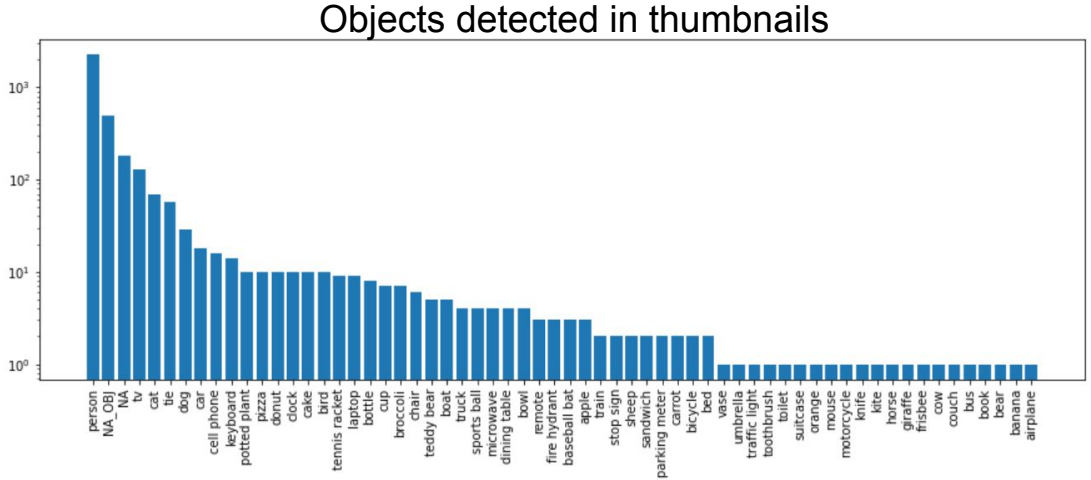
# Insights from data exploration (NLP)

Video Description Word Distribution for top 6 categories of Youtube videos in US


Video Title Word Distribution for top 6 categories of Youtube videos in US


The video description and video title word distribution for US youtube videos can be seen above. For the description, we made sure to remove stop words as well as common occurrence words like links, share, subscribe, instagram, facebook etc. The word occurrences in title and description are highly aligned to the specific categories.

# Insights from data exploration (CV)




Detected as a 'tv'

Detected as a 'keyboard'


- Top objects shown in thumbnails: person, tv, tie, cat, dog, car, cellphone, keyboard, pizza, etc.
  - tie: More likely to be detected in videos that is in a formal place, such as news or shows.
- Predicted objects might not actually be the object we thought but still give us some information
  - tv: Most of the objects detected as tv are actually some blocks containing an image.
  - keyboard: Most of the objects detected as keyboard are actually some blocks containing words or sentences.
- Notes:
  - NA_OBJ: no object detected in the thumbnail
  - NA: no thumbnail available

# Insights from data exploration

- Evening and noon/afternoon are the best times for videos to trend as most people have finished their daily activities and these videos become more noticeable.
- Most people go on youtube to enjoy themselves with entertainment-related content (music, how to and style, people and blogs, comedy).
- Most videos are published and being interacted with their audience in the warmer months (spring and summer)
- Video metrics are highly related to each other (More views results to more comment, more likes and more dislikes)
- But at the same time, the conversion rate (like_rate, dislike_rate and comment_rate) does not depend on the number of views.
- The word clouds for each category has different distribution but there is a significant difference even in the title and description of the videos. Example: Entertainment and music titles have a large occurrence of the word ‘official’ but it is not present in the description.
- A very high proportion of thumbnails have a person in the picture.

# Machine Learning techniques proposed

- Goals:
  - Classification:
    - Predict category of the video
    - Detect malicious/biased title/descriptions.
  - Regression: Predict like, dislikes, and comment counts
  - Recommendation: Suggesting video title and suggesting objects shown in the preview image.
- Feature extraction:
  - Apply pre-trained super resolution model on the thumbnail images to enhance the image quality.
  - Use pre-trained encoder models to extract features from enhanced thumbnail images.
  - Use BERT or some other pre-trained encoder models to extract feature from descriptions.
- Classification and regression:
  - Apply different classification and regression methods and use ensemble method to achieve above mentioned goals of predicting category, number of likes, views, dislikes.
  - Experiment transfer learning (train with USA data) on different countries to evaluate their like/dislike count and video category.
  - Extract samples from different countries and train the model and evaluate on different countries
- Recommendation:
  - Use deep learning models to generate suggested video title and objects for preview image.