

YouTube Trending Video Analysis 2020-2021

Kaustubh¹, Narinder Kaur²

¹UG Student, Maharaja Agrasen Institute of Technology, Dept. of IT, Delhi, India

²Professor, Maharaja Agrasen Institute of Technology, Dept. of IT, Delhi, India

Abstract - YouTube is a big multinational company under the umbrella of tech giant Google. YouTube is not only serving as an entertainment platform for the films and the television industry but it has also emerged as a learning platform for many students. Content creators on YouTube, commonly known as "YouTubers" are pushing their content every single day and hour to be relevant to their audiences. It's a known fact that the YouTube algorithm, which recommends videos to the audience searching for a particular topic or videos on the homepage is still not predictable and the YouTube development team has not made any bold claims about this. The trending section of YouTube is also one such entity whose inner working is still not out. The trending section is where some top 50-70 videos are listed as trending and it is designed in such a way that every user of this platform checks it once in a while. Some of the most popular creators were discovered via this trending section only. This analysis aims to analyze the trending section of YouTube for India and try to conclude some of the trends and observations which could be helpful for creators to push their content to more audiences.

Key Words: YouTube, YouTube trending section, Video metadata, Python, Data Analysis, Pandas library

1. INTRODUCTION

A particular video content on YouTube is not limited not to the actual video content being displayed to the user. A user visiting a YouTube video is encountered with likes, dislikes, video title, description, comments publishing date, publishing channel, number of subscribers of the channel and much more elements. All these features directly or indirectly contribute to a video being on the trending section.

This study can help in finding and comparing key aspects of a YouTube video and help budding creators, businesses to align their content to be included in the trending section in future.

2. TRENDING VIDEOS ON YOUTUBE

The YouTube "Explore" tab allows users to see what's trending in their country, as well as which events or videos are being watched the most. In the Trending section, videos with a large number of viewers who find the video interesting are displayed. Some Trending videos which are expected to be on this list include a new movie trailer, a new

song from a well-known musician or a video from a very popular content creator.

On average, every 15 minutes, the list of videos in the Trending section are updated. [1] Videos might stay in the same spot in the list, move up or down with each update. This section is supposed to be biased free and any user is not allowed pay to have their videos appear in trending results.

2. REQUIREMENT AND DATA GATHERING

Requirement gathering includes visiting the YouTube trending section and picking up random videos for potential data points. These data points decide what affects a video performance. Some of the examples include date of publishing the video, title of the video, YouTube channel type (individual or owned by company), subscribers count, trending date of the video, tags used, count of views, likes, dislikes, & comments, description of the video.

The dataset used for this analysis is a Kaggle dataset obtained from the internet. [2] The dataset is updated daily and has data for almost all the countries. We are analyzing data for Indian YouTube video data which has close to 87,000 videos metadata.

3. DATA WRANGLING

Data wrangling usually entails converting and mapping data from one raw format to another in order to facilitate data consumption. [3] The data should not have any null, special or unrecognized characters. The inferred data type should be correct. Pandas is one such library that loads the dataset and infers the datatypes. [4] The data frame loaded still needs correction.

On an abstract level, these were the steps performed on the data loaded (as Pandas Data Frame):

- Dropping null values. These are the rows whose values are null. We can see that null values are denoted by NaN. These null values are replaced with an empty string, in case of string type column and 0, in case of integer column.
- The date columns should be in datetime format used by Pandas. It helps in accessing the specific date attributes.
- All the integer data types are inferred as int64 which is not an efficient approach to store a value as

small as one digit.[5] It affects speed of execution for further data analysis. Therefore, these also need to be in the correct integer range.

4. ANALYSIS ON BASIC STATISTICS OF TRENDING SECTION VIDEOS

With the help of Pandas describe function, we can easily generate a statistical summary which includes mean, count, std deviation, percentiles, and min-max values of all the features. [6]

	categoryId	view_count	likes	dislikes	comment_count
count	87,782.0	87,782.0	87,782.0	87,782.0	87,782.0
mean	20.7	3,006,039.4	149,866.5	6,382.1	13,140.2
std	6.2	7,920,213.9	460,647.8	121,676.5	108,657.6
min	1.0	0.0	0.0	0.0	0.0
25%	20.0	419,141.2	14,818.2	526.0	619.2
50%	24.0	997,985.5	40,532.0	1,381.0	1,880.0
75%	24.0	2,479,289.5	114,137.8	3,771.0	6,182.0
max	29.0	264,407,389.0	16,115,240.0	12,341,474.0	6,738,565.0

Fig -1: Pandas Describe output for the dataset

Here are some pointers from the summary generated:

- 50% of the data shows that the trending videos had an approximate of 10 lakh views. The average number of likes of a trending video is about 1.5 lakh while the average number of dislikes is 6 thousand.
- It shows that a minimum of 0 views has gone to trending section. This is due to the fact that some of the videos had special customizations by YouTube team. Their view count was hidden from the general public and that's why the view count is zero.

5. ANALYSIS ON VIDEO TITLE

The first thing that a user checks in a video is its video tile. The title represents what content the user can expect in the video. In the trending videos for India, the title has a mix of both Hindi and English language. The titles also have numbers, special characters, punctuation marks and emojis. A video title can be analyzed in 2 different ways.

5.1 Video Title Length Analysis

We define video title length as the number of words contained by the title. We created a Python script that cleans the title and counts all the words in the title. The cleaning process includes removal of emojis, numbers, special characters, punctuation marks etc.

On getting the title length for each video, we plotted a box plot using Python's seaborn package. [7]

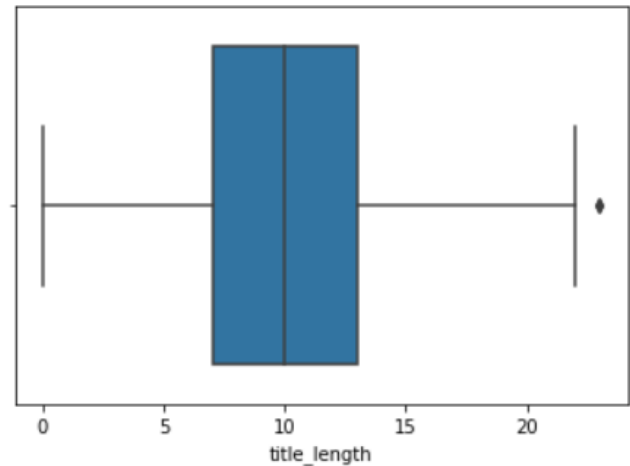


Chart-1: Box Plot for video title length

The above box plot suggests that 50% of the videos had video title length of 10 words. It also suggests the range of video titles for 50% video is from 7 to 13 words.

5.2 Video Title Words Analysis

We want to find out what words are used in a Trending video. The words used in the title directly affect if the video will be watched by the user or not. For this, we created a Python script that counts the unique words from all the video titles of the dataset.

- ```
[('video', 8231),
 ('official', 6545),
 ('the', 6507),
 ('new', 6207),
 ('song', 6042),
 ('episode', 5713),
 ('in', 4924),
 ('punjabi', 4700),
 ('th', 4629),
 ('to', 4570),
 ('ep', 4511),
 ('songs', 4285),
 ('latest', 4118),
 ('vs', 3892),
 ('with', 3651),
 ('ft', 3648),
 ('shorts', 3642),
 ('tv', 3555),
 ('my', 3482),
 ('ka', 3413),
 ('promo', 3293),
 ('of', 3103),
 ('comedy', 2974),
 ('full', 2969),
 ('a', 2850)]
```

Fig -2: List of 25 most common words

The above list can also be presented as a word cloud. A word cloud is a visual representation of words.[8] Popular words and phrases are highlighted using cloud creators depending on their frequency and importance. They give you quick and easy visual insights that can lead to more detailed analysis.



Fig -3: Most used keywords in trending videos titles

### 6. ANALYSIS ON YOUTUBE CHANNEL TYPE

YouTube allows any video, irrespective of their status whether the channel is managed by individual creator or a company/business to be in trending section. We want to check how many videos in trending section are published by a corporate company vs non company. For this purpose, we have plotted a horizontal bar graph.

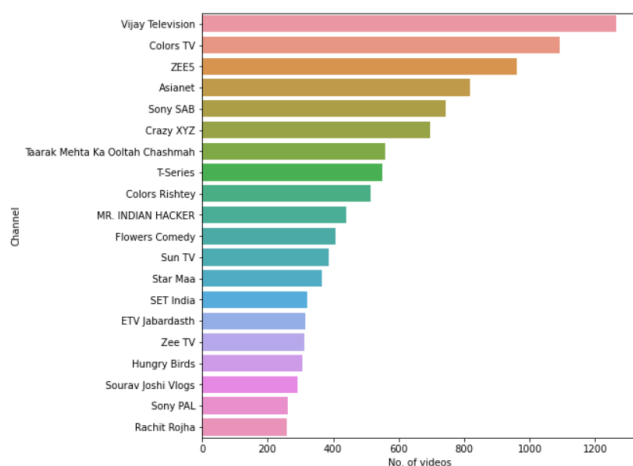


Chart-2: Horizontal Bar graph for number of trending videos per channel

On the x axis we have the number of trending videos and on the y axis we have the channel name. The bar chart tops with 1200+ videos from a corporate channel “Vijay Television” followed by “Colors TV”, “ZEE5”, “Aisanel” and “Sab TV”. Non corporate channel includes “Crazy XYZ”, “Mr. Indian Hacker”, “Saurav Joshi Vlogs etc.”

We can clearly see that although corporate channels have most amount of trending video, but the non-corporate channel are still making their way to the trending sections.

### 7. ANALYSIS ON YOUTUBE VIDEOS CATEGORIES

Whenever someone uploads a video on YouTube, it has an option to add the category for the video. Most of the users add this category so that their videos are recommended to the right type of audience. We want to see what content category gets into the trending section. We have plotted a bar graph using Python’s seaborn library. [9]

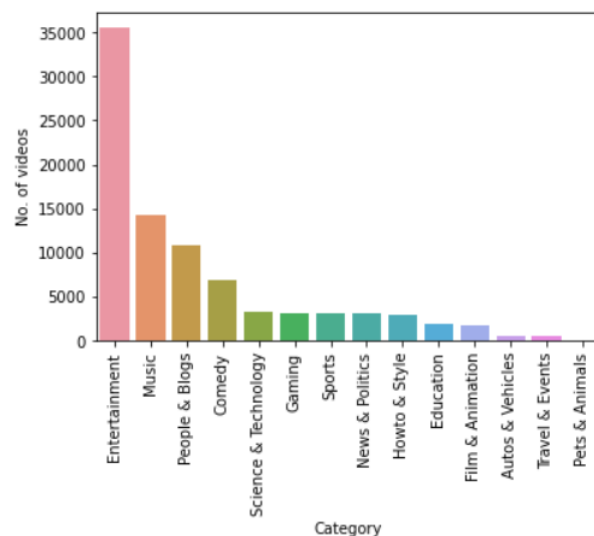


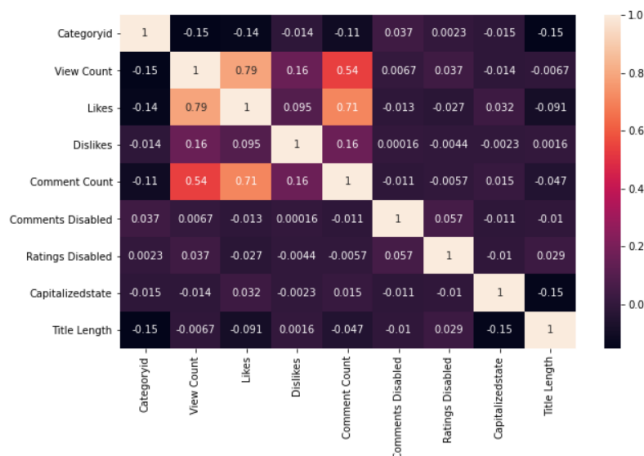
Chart-3: Bar graph for number of trending videos per category

From the above chart, it is clear that videos with category as “Entertainment”, “Music” and “People & Blogs” top the trending section. It is interesting to note that “Education” category has made it to the top 10. The rise of “Gaming” category is also notable. In 2020-21 phase, gaming creators were on a rise. Creators like “Gamer Fleet” have gained mass audience support in this phase.

### 8. CORELATIONS BETWEEN LIKES, VIEWS AND COMMENTS

The degree to which two features move in regard to each other is measured by correlation. It is expressed as a number between +1 and -1, with +1 denoting a strong positive correlation, -1 denoting a strong negative correlation, and 0 denoting no correlation.

The correlation is easy to visualize by using a heatmap. We plotted a heatmap using Python’s seaborn library. [10]



**Chart-4:** Correlations between all the numeric (and Boolean) features of the dataset represented in Heatmap

Views and Likes have a correlation of 0.79, Likes and Comments have a correlation of 0.71, and Dislikes and Comments have a correlation of 0.16. There is a positive correlation between Likes and Dislikes, Views and Comments, and Likes and Dislikes. According to results gathered by analyzing features such as ratings, just 1% of Trending Videos restricted users from responding.

### 9. ANALYSIS ON VIDEO TAGS

When a user posts a video to YouTube, they have the option of adding tags to it. Tags aren't visible on the video page, but one can see them by looking at the page's source code or using browser extensions like vidIQ.

50% of the popular videos have 19 tags, with 12% having no tags. The smallest number of tags in a trending video is 1. We have built a word cloud with the greatest number of words used in these categories using the word count package and matplotlib.



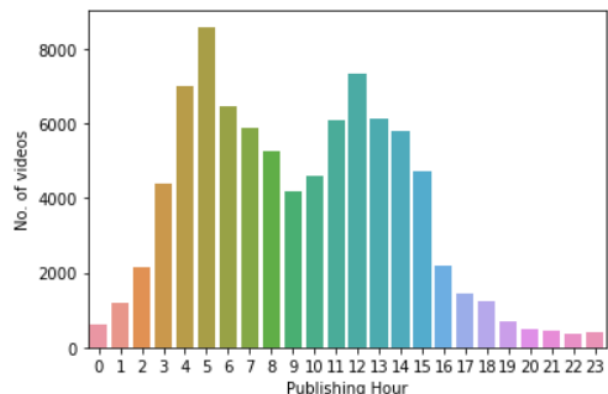
**Fig -4:** Most used tags in trending videos titles

### 10. ANALYSIS ON DATE OF VIDEO PUBLISHING

We can figure out the optimal time for a trending video to be uploaded by combining the publishing time for each video with the number of videos for each time slot. As a result, we can forecast whether a video in a specific time slot with a specific set of attributes will trend or not using algorithms like regression.

The following are some summaries based on day and time analysis.

- Midweek afternoons and early morning 5 am are the best day and time to publish video to YouTube.
- After analyzing data from 263 billion views over 2800 channels, we discovered that the optimal time to submit videos are on Thursday and Friday from 11 to 4 p.m.
- September-October came out to be the biggest contributor of trending videos in 2021 as compared to May with lowest contribution.



**Chart-5:** Bimodal bar graph showing Publishing hours of trending videos

### 11. KEY RESULTS OF ANALYSIS

**Table -1:** Analysis for year 2020-2021

| Characteristics                                 | 2020-2021                              |
|-------------------------------------------------|----------------------------------------|
| Dataset                                         | Video metadata for about 87,000 videos |
| Video title length (in terms of how many words) | In the range of 7 to 13 words          |
| Largest Category                                | Entertainment                          |
| Largest Channel                                 | Vijay Television                       |

|                                           |                                                                                                                                        |
|-------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------|
| Common Words in Video titles              | "Video", "Official", "New", "Song"                                                                                                     |
| Publishing Hours                          | 4 to 5 am or 11 to 4 pm.                                                                                                               |
| Video on trending list                    | Video can remain on the trending list for more than one day                                                                            |
| Correlation of video features             | Likes and views are highly positively correlated.                                                                                      |
| Views                                     | About 50% of the trending videos have 1 million plus views. Some special videos with zero views also appeared in the trending section. |
| Video with Maximum views                  | "BTS 'Butter' Official MV" 264 million views                                                                                           |
| Largest Channel (in terms of subscribers) | T-Series                                                                                                                               |

[7] "Seaborn Boxplot"  
<https://seaborn.pydata.org/generated/seaborn.boxplot.html>

[8] "What is a Word Cloud"  
<https://monkeylearn.com/word-cloud/>

[9] "Seaborn Bar plot"  
<https://www.geeksforgeeks.org/seaborn-barplot-method-in-python/>

[10] "Seaborn Heatmap - A Comprehensive Guide"  
<https://www.geeksforgeeks.org/seaborn-heatmap-a-comprehensive-guide/>

## 12. Conclusion

Our findings for measuring, assessing, and comparing essential characteristics of YouTube popular videos were reported in this study. Knowing the optimal time to publish a video to YouTube isn't enough to get millions of views and make your video popular. Other elements to consider are good titles, good thumbnails, video SEO, proper tagging, and the number of subscribers, all of which are important in driving views for your material. Understanding these statistics will aid YouTube in not only developing better video processing algorithms, but also in making judgments for individual youtubers.

## REFERENCES

[1] "How to Use YouTube to Grow Your Channel"  
<https://keywordtool.io/blog/youtube-trends/>

[2] "YouTube Trending Video Dataset (updated daily)"  
<https://www.kaggle.com/rsrishav/youtube-trending-video-dataset>

[3] "What is Data Wrangling?"  
<https://www.altair.com/what-is-data-wrangling/>

[4] "What is Pandas in Python?"  
<https://www.activestate.com/resources/quick-reads/what-is-pandas-in-python-everything-you-need-to-know/>

[5] "Optimizing the size of a Pandas data frame for low memory environment"  
<https://vincentteyssier.medium.com/optimizing-the-size-of-a-pandas-dataframe-for-low-memory-environment-5f07db3d72e>

[6] "Pandas Describe"  
<https://www.machinelearningplus.com/pandas/pandas-describe/>