Statistical Measures for Trending Video Analysis using knowledge pre-processing and preliminary knowledge Analysis

Rahul Pradhan, Akansha Saxena, Divyanshi Bansal and Ananya Jain
GLA University, Mathura
akansha.saxena0211@gmail.com

Abstract. YouTube has an algorithm that decides which videos will appear in the Trending Videos List. Some are more likely to be on the trending list but some videos which were never estimated to be on-trend are seen, it targets viewers’ attention over a short span. On Youtube, trending video lists are not personalized according to a person’s views, likes, or dislikes just like what the recommendation system does, rather its algorithm targets a collective group of people, the same Trending video page is displayed to everyone in a country. In India, the Trending page of: Like various recommendation systems, YouTube also gives you suggestions on your Home page based on your previously watched content. Recommendation systems are so common that we are using it without even knowing it, it makes the user experience much better. Common Examples are Amazon, NetFlix, and now Spotify uses it to suggest to us the music we might love to listen to. The trending videos page updates itself every few minutes (15 minutes). After every update, videos may move up, down, or stay in the same position in the list. Here in this Paper, we used statistical skills to analyze what are the reasons for a video to be on trends. In this paper, we have worked on a dataset for the year 2020. For data collection, we used a Web Scraper script, that’s not part of our work. Based on EDA we can say what is the exact reason for a video to appear on the trending page.

1. Introduction
There is no other platform as popular as YouTube when it comes to video content. The video uploading rate is very high and billions of views are there in a single day. There are times when you search for motivational videos or some videos to save your life during exams, talking about tonnes of educational content available.

Try to keep some points in your mind while uploading your very own video:
• Videos must be appealing to a wide range of viewers
• You must ensure that videos are not at all misleading
• Capturing the breadth of what's happening in the world
• Work on unique ideas that might bring more and more viewers to your video.
• A video might appear on a trending page due to some of these reasons:
  o View count
  o Rapid growth in views with time
  o Where views are coming from (including outside of YouTube like Google search, Facebook, etc.)
The exponential increase in the amount of data that is generated every year is demanding us to look at it as data is becoming more and more critical. The analysis includes simple queries, statistical analysis, and data mining. Having so much content available, social media streams what people are paying attention to and what is the hype about certain topics. It becomes a need to have data analysis’s conclusions. Correctly utilizing trending topics requires a better understanding of their various characteristics in different social media streams and it requires statistical skills to draw some conclusions. As part of our Project, we have collected datasets through YouTube data API which provides a systematic collection of public standard feeds and statistics related to videos and users. We aim to produce a scientific knowledge preprocessing analysis operating solely with the dataset INDIA Videos. This step is important for all data processing exercises and that we wish to emphasize it. We will first grasp key knowledge attributes, like missing values, distinctive counts, outliers, and time-series trends.

This Paper aims to function as a help to find trends in the video which will help a channel to grow and will get to know about what kind of videos are coming in trends. The focus is only on the INDIA videos dataset that isn't too huge by big-data standards. This knowledge set contains solely YouTube data like video, image, audio, or giant text documents. Still, we are going to proceed with knowledge preprocessing and preliminary knowledge Analysis (EDA) as if this were a very huge dataset, using techniques that might be utilized in rather more difficult knowledge manning exercises.[1]

The data used in this analysis were retrieved using the Youtube API. A script was to fetch data from YouTube data API about trending videos of each day. The script in which the youtube API key is passed after making an account at google console for YouTube and then processes the data received from YouTube and stores them in text files. The scraping script is not part of our work, only worked on the EDA part [2].

![Figure 1. The workflow of a data analysis project](image)

2. Related Work

(Barjasteh et al., 2014) [3] mentioned that there is no such statistical analysis on the trends of YouTube videos. This statistical analysis hasn't received the attention which it should have received seeing the significance of this analysis. With the years passing by YouTube has become the best place for advertising your brand, people might show their response be it positive or negative, or maybe neutral. They analyzed and evaluated 4,000 trending videos that were collected for about 2 months using YouTube API. Dataset itself describes the number of views, comments, time duration, its category. It includes a comparison between trending and non-trending videos, from which they were able to draw some statistical attributes out of trending videos. Analyzed the profile of uploaders. It might not seem so obvious but it can influence the video. Basically what they analyzed were the subscriber’s count and other uploaders’ stats. They also analyzed the directional-relationship between the trending videos and the trending videos that became popular. They used Granger Causality(GC) with significance testing to measure the directional relationship. [1]
(Bärtl., 2018) [4] aimed at how video consumption evolved over the course of time. Through the results of an analysis, it was evident that older videos have a higher probability to gain large viewership than the young channels, it all depends on the topic they select for making their video. Since this analysis was done on a dataset of 10 years, as a result, they were able to observe the gradual growth of various channels. So what they were focusing on was how YouTube evolved in terms of channels, uploads, and views. Identified what were the factors that were responsible for channels’ growth in terms of viewers’, subscribers’ count. [4]

YouTube videos are uploading rapidly in real-time. Users can share, comment, and rate (like or dislike) videos [6]. (Bhuiyan et al., 2017) [5] worked on NLP processing based sentiment analysis on comments by the users., which helps in finding out the emotion of the video. To date, YouTube is the best platform for the compilation of video content on the web. The versatility and variety of content attract users across the globe. Sometimes it also happens that we get nothing we are looking for. Happens mostly when we are searching for a particular movie instead what we get is some movie’s cut scenes or its trailer. This kind of scenario occurs frequently due to likes that video has got. [3] With all these things happening on a platform, there is some kind of interest in analyzing the rich content shared on YouTube. It has been seen that comments might be a good source to perceive the video content i.e its quality, the relevance of the topic, and gives us the idea of what kind of message the video is giving. [3]

![Figure 2. Workflow for sentiment analysis](image)

(Stefan Siersdorfer et al., 2010) did an in-depth study of comment rating behavior of around 6.1 million comments on 67290 videos on YouTube. They also analyzed what were the exact dependencies between topics, views, comments. They also studied the sentiment of comment and its impact on the video rating, for this they used SentiWordNet thesaurus. They created their test collection by formulating queries from Google’s Zeitgeist archive from 2001 to 2007 For each video around 500 comments were gathered and they all were analyzed for comment rating. [6]

There are various techniques for sentiment analysis which are performed on user comments and for this SentiWordNet is used majorly. In sentiment analysis we deal with mainly 3 issues: Classification of events, Sentiment polarity detection, and prediction of youtube comments. Detection of sentiment polarity could be Lexicon based, negation based or SMAPD. Evaluation is done on the basis of precision.
and recall. If you make a search on your favourite search engine, what it does is it returns 30 pages (say) and only 20 pages are relevant. Then its precision is 20/30. After that classification is done using SVM. Classification is done on a dataset of user comments that is divided in two parts i.e Test data and Training data.[7]

![Sentiments](image)

**Figure 3. Sentiments**

3. Proposed Approach
There are a series of steps involved for a conclusion before getting results of the analysis on data:
- Data Collection
- Data Preprocessing
- Data Visualization
- Build Model

The analysis is done on the basis of likes, dislikes, title, description, comments.
Based on this we analyzed what was the reason for a video to be trending. We extracted data for performing data analysis on YouTube videos to get into the latest trends. Since YouTube data grows rapidly with a great speed, that is why there is a huge demand to store and process it.

Before data insights, we would remove missing values, distinctive counts, outliers, and time-series trends that are a part of the data cleaning process.[8]

3.1 Problem Definition
This paper is based on the analysis of Trending Videos on YouTube. Nowadays, every 10th person has a YouTube channel and is uploading video content. Some videos become popular which leads to the growth of a channel but for some channels, it takes too long to grow, it all depends on what topic that channel is covering through its videos.

There are various types of content available on YouTube:
- Entertainment
- Sports
- Music, Genres, Albums
- Daily Soap
- Trailers
- Workout videos of fitness experts
- Cooking Videos
- Educational videos
- Romantic videos
- Spiritual videos

And it becomes interesting as well as significant enough to explore the insights that why a video is trending, what is it depending on; whether it is the category or the particular title which is catching users’ attention. We want to dig deep to analyze the relationship between various characteristics of trending videos. Some videos are controversial due to the content it has. Like a few days back in 2020 there were lots and lots of videos on the Sushant Singh Rajput case. Many YouTubers made videos on it as the title of their videos. We want to analyze how cultural divergence is affecting likes, comments.

3.2 Dataset Description

We created a dataset named “Trending Video Analysis” which was built using YouTube API Services, for this, we have merged several CSV files which were having data of a couple of weeks on trending videos from YouTube for India. Dataset is having more than 1000 observations, having information such as video title, category, publish time, view count, likes, dislikes, comments, description, channel id, video id, etc. Before working on this we had a few questions like what are the characteristics which trending videos possess. We measured user engagement through these characteristics.

3.3 Plan

The analysis was performed using a Jupyter Notebook
We examined view counts on a video. We also examined likes and dislikes for a video that is trending to get to know what is the viewers’ attitude towards that type of video. We also get to know what were the common words found in the title of a trending video like: “official”, “trailer”, “music”, etc.

3.4 Visualization

We used: Seaborn, Numpy, Pandas. We also used NLP Libraries. We performed this analysis using Python and Python libraries like Pandas, Matplotlib, NLTK, WordCloud.

4. Analysis Results

Through the results of the analysis on our dataset, we are able to see that the trending video having the highest view count is by BTS band names ‘Dynamite’OfficialMV. It has around 16124714 likes and 757828 dislikes.

Table below is the information about the trending video having the highest view.

| Video Id          | gdZLi9oWNZg               |
|-------------------|---------------------------|
| Title             | Dynamite Official MV      |
| Channel Title     | Big Hit Labels            |
| Likes             | 16124714                  |
| Dislikes          | 757828                    |
| View Count        | 254687524                 |
| Comments Count    | 6255536                   |
4.1 Most common words in the trending videos

While analyzing the dataset we also found what were the most common words in the trending videos.

![Graph showing commonly occurring words in trending videos](image)

**Figure 4.** Commonly Occurring Words in Trending Videos

![WordCloud for commonly occurring words in trending videos](image)

**Figure 5.** WordCloud for commonly occurring words in trending videos

Figure above is a representation of Fig 4 in Word Cloud which represents the words on X axis (Commonly occurring words in Trending Videos)

4.2 Channel titles

We also got to know through our analysis that what are most Channel titles of trending videos

Those were:

- Technical Guruji
• ZEE5
• Taarak Mehta Ka Ooltah Chashmah
• Speed Records
• SAB TV
• Desi Music Factory
• Colors TV
• SET India
• Anysaa

There were some more of them, but here we listed only those which we all are aware of, to just get a glimpse of what are the channel titles which became trending so that you instantaneously get to know what would be the type of content. You can see it through the graph below.

**Figure 6.** Common Channel titles

Image above is small so the channel titles on X axis are mentioned in the bulleted list.
4.3 Views on Trending Videos

We also get to know what were the Views on Trending Videos

![Word cloud for common channel title](image)

**Figure 7.** Word cloud for common channel title

![Graph showing views on trending videos](image)

**Figure 8.** Views on Trending Videos

Through the figure above we can observe that the majority of videos have views less than 5 Million.

4.4 Videos with Less than 5 Million views
Figure 9. Views of Trending Videos

(In Figure 9 data has been selected in order to have videos lying in the category of views less than 5 million)

From the figure above we can see that most of the videos have less than 1 Million views

4.5 Likes on Trending Videos

Figure 10. Likes on Videos

From the above figure, we can get to know that most of the trending videos have liked by less than 5 Million.

4.6 Likes on Trending Videos Below 20,000
5. Conclusion

From the results of our analysis, we can conclude that we found some obvious channels on the trending list. Those were: Technical Guruji, ZEE5, Tarak Mehta Ka Ooltah Chashmah, Speed Records, SAB TV, DesiMusic Factory. We found that most of the videos have less than 1 Million views. Common words found in the title of the trending videos were: “Official”, “Trailer”, “new”, “Punjabi”, “song”. The trending video having the highest view count is by a BTS band named “Dynamite'OfficialMV”.

References

[1] Barjasteh, I., Liu, Y., & Radha, H. (2014). Trending videos: Measurement and analysis. Retrieved from http://arxiv.org/abs/1409.7733
[2] Amudha, S., Niveditha, V. R., Kumar, P. R., Revathi, M., & Radha, S. (n.d.). Youtube Trending Video Metadata Analysis Using Machine Learning.
[3] Bhuiyan, H., Ara, J., Bardhan, R., & Islam, M. R. (2017). Retrieving YouTube video by sentiment analysis on user comments. 2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA). IEEE.
[4] Bärtl, M. (2018) ‘YouTube channels, uploads, and views: A statistical analysis of the past 10 years’, Convergence, 24(1), pp. 16–32. DOI: 10.1177/1354856517736979.
[5] Stefan Siersdorfer, Sergiu Chelaru, Wolfgang Nejdl, and Jose San Pedro. ( 2010). How useful are your comments? analyzing and predicting youtube comments and comment ratings. Association for Computing Machinery, New York, NY, USA, 891–900. DOI:https://doi.org/10.1145/1772690.1772781
[6] Mahrishi, M., Hiran, K. K., Meena, G., & Sharma, P. (2020). Machine Learning and Deep Learning in Real-Time Applications. IGI Global.
[7] Siersdorfer, S., Chelaru, S., Nejdl, W., & San Pedro, J. (2010). How useful are your comments?: Analyzing and predicting youtube comments and comment ratings. Proceedings of the 19th International Conference on World Wide Web - WWW ’10. New York, New York, USA: ACM Press.
[8] Mohana Prabha, B. M., & Ramya, R. P. (Eds.). (n.d.). Predicting the Popularity of Trending Videos in Youtube Using Sentiment analysis.