YouTube Trending Videos Analysis Using Hadoop

Authors: Sapan Shah, Snehil Sarkar, Sai Sridhar Karri, Kaushik N Adhikari

Department of Information Systems, California State University, Los Angeles
CIS 5200 System Analysis and Design
sshah82@calstatela.edu, ssarkar4@calstatela.edu, skarri2@calstatela.edu, kadhika3@calstatela.edu

Abstract: This big dataset is all about collection of YouTube trending videos data. Data is captured from different countries and updated on daily basis which includes title, channel, time, tags, views, likes, dislikes etc. By using Hadoop & Hive, certain amount of data is being performed and then analysis & visualize all through Excel 3D-map & Tableau.

The focus on this data analysis is mainly on categorizing videos by different factors and exploring sentiments in variety of forms. By analyzing registered counted views, likes, dislikes and comments, current trends can be figured out. Combining data of different regions of world can generate the comparative analysis like how videos are shared across the nations and gained popularity.

1. Introduction

YouTube is one of the major online social media platforms providing online videos sharing. This is a subsidiary company owned by Google. It was launched in USA in February 2005 and now the second most visited website in the world. Every minute, more than 500 hours of content is uploaded and around 2.5 billion users have accessed this website every month.

This escalated popularity has drawn attention significantly and now become more important to analyze its dataset thoroughly so the viewers and content creators can take the advantage the most. Additionally, this dataset is very large in terms of diversity so some of our analysis is solely focused on one country, but similar methods can be used to examine data of any other countries to achieve the same solutions.

By analyzing this dataset through Hadoop and Hive, we can gain valuable insights of viewers sentiments and different aspects responsible for the popularity of trending videos. Our objective is to find top 10 trending videos worldwide[1]. We have also figured out the most trending videos of individual countries based on viewers engagement score. Analysis conducted year-on-year basis to find the most viewed YouTube channel in USA and identified the content preferences and marketing opportunities. We have also extracted the most frequent tags used while searching the videos to understand the prevalent themes or subjects of YouTube videos in INDIA. Part of analysis is to examine various categories of videos that has been viewed in all the countries and identify public interest on each year.

2. Related Work

As YouTube is one of the major online platforms in social media world so many researchers have analyzed trending videos dataset. One such research paper was submitted on International Research Journal of Engineering and Technology (IRJET) in August 2020 by S Gayakwad, D Mane and R Patankar [1]. They mainly focused on studying the difference between trending and non-trending videos, analyzing youtubers who published most trending videos, other basic statistics, video length, category and published time by using algorithm like linear regression, other machine learning models and python libraries like panda and matplot.

In 2021, another paper was published by Johanes F Andry & other faculty members from Jakarta, Indonesia [2] on Algorithm of YouTube trending videos. By using RapidMiner software for datamining, dataset was analyzed and come up with the conclusion that there are two main factors that affects algorithm, engagement of viewers and metadata. To achieve results, authors has analyzed dataset with classification, association and clustering method.

G Mohana Prabha, B Madhumitha published paper on Predicting Popularity of YouTube trending videos in 2019 [3]. They have used Support Vector Machine (SVM) classifier for algorithm and sentimental analysis with accuracy.

Our research analysis is different than all mentioned above in terms of methodology and diversified data inclusion. We have considered the worldwide data in comparison to make viewers expand their preferences beyond the local region. We have also emphasized on visualization techniques like Excel 3D map & Tableau for better understanding of our findings.

3. Specification

The YouTube trending video dataset consists of 11 individual files and each file represents data of unique country. The countries are South Korea, Japan, Brazil, Great Britain, Germany, Canada, France, Russia, Mexico, India and USA. It contains the trending data from the Year 2020 to 2023. It is continuously updated and fed through YouTube API. The dataset comprises of detailed information in the form of video id, title, published date, channel id, channel title, category id, trending date, tags, view count, likes, dislikes, comments, thumbnail and description.

[1] Data from Mexico, Germany, France and Great Britain was not used.

Below Table 1 shows files and size of the files of dataset.

Table 1 Data Specification

Total Dataset Size	3.76 GB
Dataset Size Used	2.43 GB
Total Nos. of Files	11
Nos. of Files Used	7
Content Format	CSV
Countries Included	South Korea, Japan, India, Brazil, Canada, Russia & USA

Below Table 2 shows the specification of Hadoop cluster.

Table 2 H/W Specification

Cluster Version	Hadoop 3.1.2
Number of Nodes	5 (2 master, 3 worker)
Memory Used	480.15 GB
CPU Speed	1995.312 MHZ

4. Implementation Flowchart

Initially the dataset of YouTube trending videos was downloaded from Kaggle[2] online platform where it has been continuously updated from YouTube API and this dataset comprises of all information regarding videos from year 2020 to 2023 in the form of csv files.

Figure 1 below shows the flowchart of the complete data analysis. CSV files of dataset first downloaded to local system and then uploaded to Hadoop file system. Later HiveQL queries were performed to create tables' schema, clean data and export the results. Once the required results were achieved, files sent to the local system. For visualization, we used Excel 3D map and Tableau.

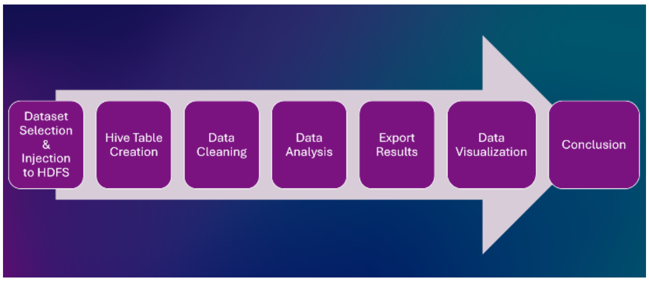


Figure 1 – Implementation of Flowchart

5. Data Cleaning

As the raw dataset was uploaded in Hadoop data file system, then the files were unzipped. We have separate files for each country, so all files were uploaded separately and loaded into tables using Beeline Client. For analysis, we need the columns like video id, title, published date, channel id, channel title, category, trending date, tags, view count, likes, dislike and comments so remaining columns were not considered during making of cleaned table. Also, files have many rows with NULL values that need to be removed as well.

6. Analysis and Visualization

Once all available data is cleaned in the form of table, HIVE queries are performed. To find the distribution of engagement of top 10 trending videos, all the 7 countries of data is taken into consideration. Engagement score is based on all the viewers' sentiments such as likes, dislikes, number of views and comments. Upon successfully run a query, results are fetched to the Linux and then uploaded to the local system.

Using Excel, distribution of engagement (Figure 2) of trending videos is visualized. Top engagement score is come up here with 15% among all other videos. These findings are the result of considering all last four years.

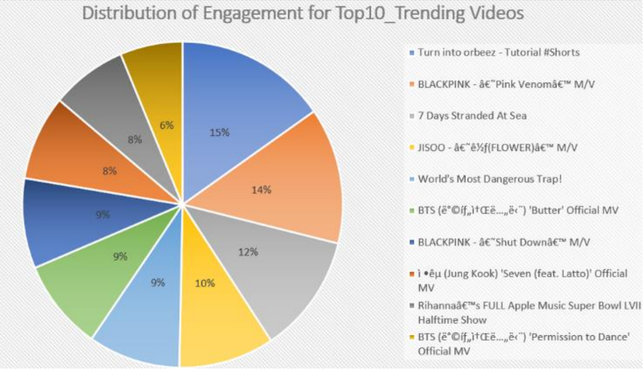


Figure 2 - Top trending videos worldwide

Similarly HIVE queries can produce the results of trending videos for individual country and using Excel, visualization through world map (Figure 3) can be possible. It can be figured out here that Brazil, Russia and South Korea share same most trending video while on other end, Canada and India are with the same preference.

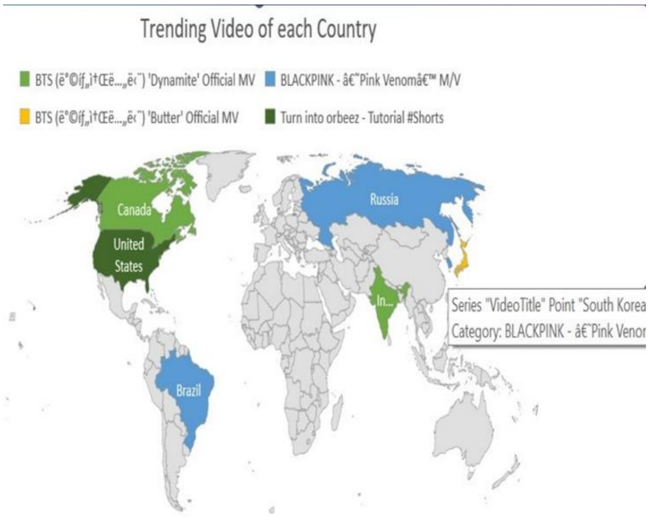

Trending Video of each Country
BTS (ë°©íƒ„ì†Œë…„ë‹¨) 'Dynamite' Official MV
BLACKPINK - â€˜Pink Venomâ€™ M/V
BTS (ë°©íƒ„ì†Œë…„ë‹¨) 'Butter' Official MV
Turn into orbeez - Tutorial #Shorts
Canada
United States
Russia
In...
Brazil
Series "VideoTitle" Point "South Korea
Category: BLACKPINK - â€˜Pink Venor

Figure 3 – Trending videos of each country

----

[2] Kaggle link: https://www.kaggle.com/datasets/rsrishav/youtube-trending-video-dataset/data

6.1 Visualization with Tableau

Every country has their own spectrum of most viewed YouTube channels. Here we consider USA to analyze their channel preferences in last all 4 years through Tableau and the most viewed channel is come up with 18.9 billion views.

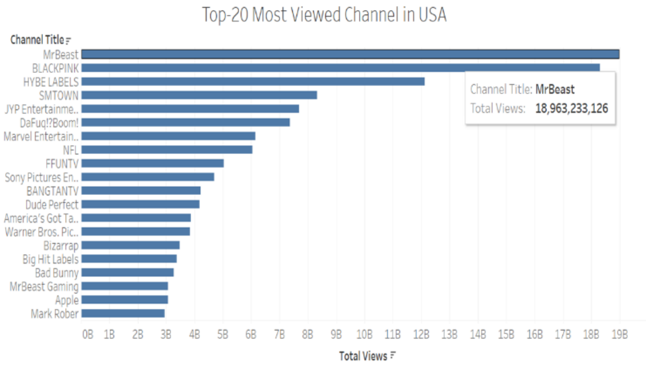


Figure 4 – Top 20 Most viewed channels in USA

When results are visualized on year basis, all different channels are populated as the most watched category. Sharp hike can be seen in number of viewers (approx..3 billion to 13 billion) from year 2020 to 2023 which shows incredible growth in YouTube popularity in USA.

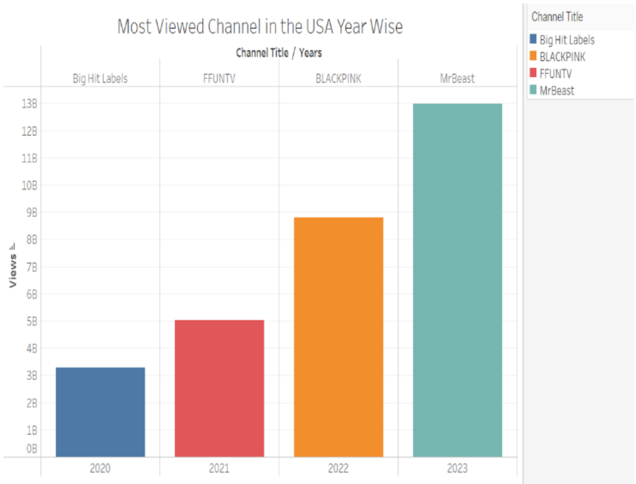


Figure 5 – Most viewed channel in USA each year

Another important factor to analyze dataset is looking upon the most frequent search words. This gives valuable insights of viewers sentiments and interests. Content creators can utilize this data for their future videos by understanding viewers expectations.

To analyze top 20 most searched words, we have used the dataset of INDIA to find the results. New, Song and Comedy are the most frequent used words out of 20 so it can be estimated that viewers of INDIA are more inclined towards latest videos, musical videos and comedy contents videos.

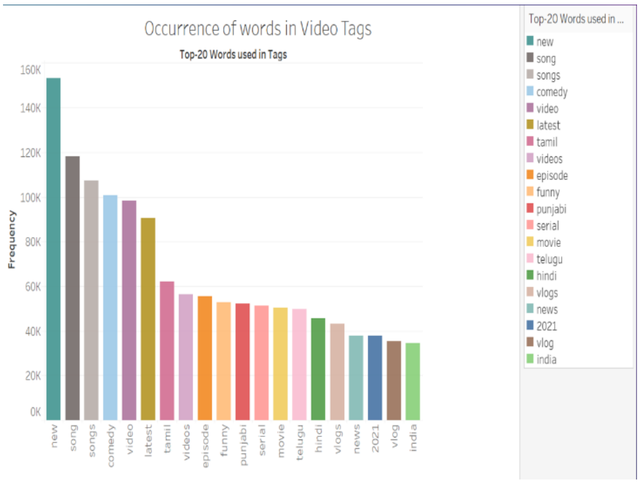


Figure 6 – Top 20 Most searched words in INDIA

The word 'new' is populated nearly 155000 times in last 4 years. Below figure 7 shows all top 20 most searched words in INDIA in the form of word cloud.

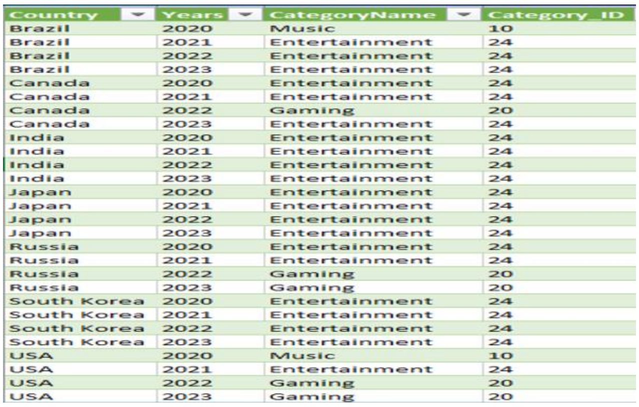
Figure 7 – Word cloud of most searched words in INDIA

6.2 Visualization with Excel- 3D map

In this dataset, all the videos are categorized with the unique genre name and category id number. Generating yearly data of all countries by their genre preferences can help to understand the distinct viewers expectations. It can also be possible to figure out some certain patterns that helps many content creators and comparing these findings with all countries provides deeper understanding of global interests.

Excel 3D map or Power map is the excellent tool to create the dynamic and interactive geographical visualization. It helps to identify trends, patterns and insights related to geographical locations.

All the YouTube videos are divided into 30 different categories but only 3 categories are the most watched around the world. Entertainment category includes videos related to award shows, cooking shows, movie reviews, celebrity interviews etc.



Country	Years	CategoryName	Category_ID
Brazil	2020	Music	10
Brazil	2021	Entertainment	24
Brazil	2022	Entertainment	24
Brazil	2023	Entertainment	24
Canada	2020	Entertainment	24
Canada	2021	Entertainment	24
Canada	2022	Gaming	20
Canada	2023	Entertainment	24
India	2020	Entertainment	24
India	2021	Entertainment	24
India	2022	Entertainment	24
India	2023	Entertainment	24
Japan	2020	Entertainment	24
Japan	2021	Entertainment	24
Japan	2022	Entertainment	24
Japan	2023	Entertainment	24
Russia	2020	Entertainment	24
Russia	2021	Entertainment	24
Russia	2022	Gaming	20
Russia	2023	Gaming	20
South Korea	2020	Entertainment	24
South Korea	2021	Entertainment	24
South Korea	2022	Entertainment	24
South Korea	2023	Entertainment	24
USA	2020	Music	10
USA	2021	Entertainment	24
USA	2022	Gaming	20
USA	2023	Gaming	20

Figure 8 – Worldwide most watched category

Below figures are showing the most watched categories of videos, visualized to their respective geographical locations with last four years of results.

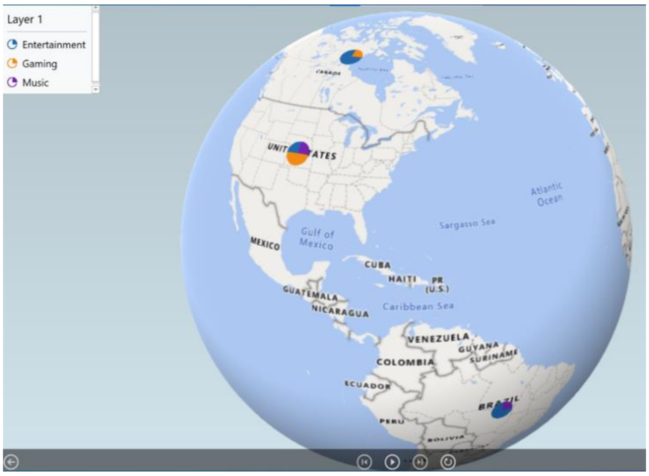


Figure 9- Category distribution in USA, Canada & Brazil

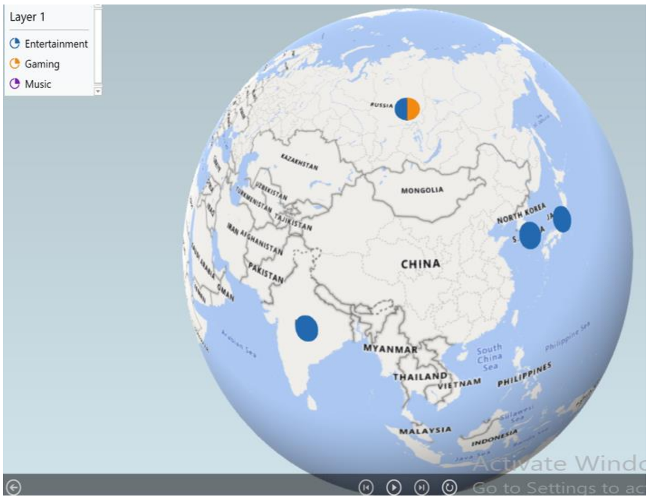


Figure 10 – Category distribution in India, Russia, South Korea & Japan

For more information, dashboards and code visit project's GitHub link[3]

7. Conclusion

In conclusion, this study has explored the valuable insights about the viewers behavior, preferences and certain patterns. By analyzing, comparing and visualizing dataset, followings are the final outcomes of our research:

- Most trending videos worldwide in last 4 years – 'Turn into Orbeez – Tutorial.'
- Most of the trending videos worldwide are originated from South Korea, such as BTS – 15%, Blackpink- 23% and Jisoo & Jung Kook- 18%
- Most viewed channel in USA in last 4 years – 'MrBeast' (18.9 billion)
- In USA, most watched channel from 2020-2023 Big Hit Labels, FFunTv, Blackpink & MrBeast respectively.
- New, Song, Comedy, Video etc. are the most frequent searched words in last 4 years in INDIA.
- In last 4 years, all countries are mostly inclined towards 'Entertainment' category of their most watched videos.

This paper certainly helps viewers to expand their horizon of preferences from local region to worldwide. Content creators can utilize this data for their better future creations and YouTube itself can develop better algorithm to process videos.

References

[1] S Gayakwad, R Patankar, D Mane, "Analysis on YouTube Trending Videos", IRJET, Vol 7, Issue 8 Aug 2020 https://www.irjet.net/archives/V7/i8/IRJET-V7I8732.pdf

[2] Johanes F Andry & others, "Algorithm of Trending YouTube videos", ICoDSE, Indonesia 2021 https://www.researchgate.net/profile/Johanes-Andry/publication/356959120_Algorithm_of_Trending_Videos_on_YouTube_Analysis_using_Classification_Association_and_Clustering/links/61b400bc4b318a6970d1d135/Algorithm-of-Trending-Videos-on-YouTube-Analysis-using-Classification-Association-and-Clustering.pdf

[3] G Mohana Prabha, B Madhumitha, "Predicting Popularity of Trending Videos in YouTube using Sentimental Analysis", IJITEE, Vol-8, Issue-6S3, April 2019. https://www.ijitee.org/wp-content/uploads/papers/v8i6s3/F10430486S319.pdf

----

[3] Git Hub URL: https://github.com/ssarkar4/YouTube_TrendingVideos_DataSet