CS661 Project : Analysis of Trending Youtube video data

Rajarshi Dutta 200762
Sourit Saha 200998
Armeet Luthra 200185
Ritam Jana 200762
Ojsi Goel 200653
Ishaan Maheshwari 200454
Jatin Chauhan 200469
Yash Goel 201142
Sushmita 201027

Abstract

This project delves into the analysis of YouTube's trending video data to uncover the dynamics behind video popularity and viewer engagement. Utilizing a dataset comprising video titles, channels, views, likes, dislikes, comment counts, and categories, we apply a series of data preprocessing, statistical analysis, and machine learning techniques. Our objective is to identify trends, sentiment patterns, and predictors of content virality. We leverage Python and its libraries for data manipulation and visualization, culminating in an interactive dashboard developed with Dash and Plotly. This visualization tool offers insights into the factors driving video success on YouTube, providing valuable information for content creators and marketers. The project highlights the importance of data visualization in making complex information accessible and actionable. The link to our repository is provided here: https://github.com/Rajarshi1001/CS661_Project

1 Introduction

In the digital era, **YouTube** has emerged as a dominant platform for content creation and consumption, influencing public opinion, entertainment, and even marketing strategies. The vast amount of data generated by YouTube provides an invaluable resource for understanding contemporary digital culture, user behavior, and content trends.

This analysis focuses on the YouTube dataset available on Kaggle, which comprises a wealth of information on trending YouTube videos. The dataset contains several features including video IDs, trending dates, titles, channel titles, categories, publish times, and various engagement metrics like views, likes, dislikes, and comment counts. This data offers a snapshot of what content resonates with the global YouTube audience, presenting an opportunity to glean insights into viewer preferences and content performance.

Our primary objectives in this analysis are to:

1. Understand the characteristics of trending YouTube videos and identify patterns and trends in viewer engagement
2. Exploring the influence of various factors such as video category, publish time, and content type on video's popularity
3. Employing a range of data processing techniques including cleaning, transformation, and exploratory data analysis. Through this meticulous process, we intend to ensure the integrity and quality of the data, facilitating a robust and meaningful analysis.

The insights derived from this analysis hold immense value for content creators, marketers, sociologists, and platform developers, offering a lens into the rapidly evolving landscape of online video content. By understanding the dynamics of trending videos on YouTube, stakeholders can make informed decisions, tailor their strategies, and perhaps even anticipate the next viral sensation.

2 Dataset Description

The chosen dataset is a secondary dataset sourced from Kaggle. It is about YouTube (the world-famous video-sharing website) which maintains a list of the top trending videos on the platform. This dataset is a daily record of the top trending YouTube videos. It was collected originally using the YouTube API. The dataset comprises several CSV files, each representing a specific country's YouTube video statistics. The countries included cover a diverse range of geographic areas, providing a broad perspective on YouTube content consumption globally. Key features of the dataset include:

1. **Video ID and Title**: Unique identifiers for each video along with their titles.

2. **Channel Title**: The name of the channel that uploaded the video.
3. **Channel Title**: The name of the channel that uploaded the video.
4. **Category ID**: An identifier for the category of the video.
5. **Publish Time**: The date and time when the video was uploaded to YouTube.
6. **Tags**: A list of tags associated with the video, which are keywords or phrases describing the video content.
7. **Views**: The number of times the video was viewed.
8. **Comment Count**: The number of comments on the video
9. **Thumbnail Link**: The URL of the video's thumbnail image.
10. **Comments Disabled and Ratings Disabled**: Boolean fields indicating whether comments or ratings are disabled for the video.
11. **Video Error or Removed**: A field indicating whether the video had an error or was removed.
12. **Description**: The video's description.

Since, there are multiple CSV files in the dataset folder, the region files were merged along with an additional column indicating the region and below is the bar plot (1) illustrating the number of videos per region. It is observed that Russia contributes to the largest umber of unique videos followed by Mexico with United Kingdom being the lowest contributor in the overall dataset.

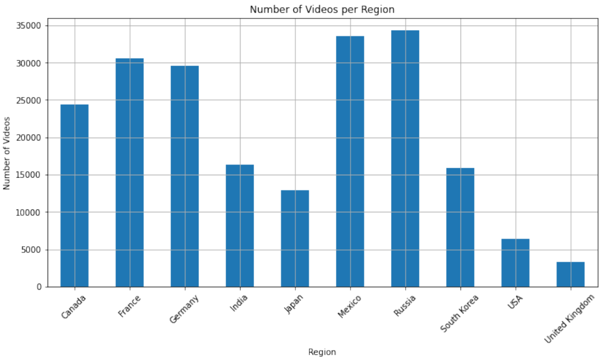

Figure 1: Number of Videos vs Region

3 Tasks and Methodology

As for the initial steps to be performed for this dataset, these include **Data Cleaning**, **Data Transformations**, **Feature Extractions**, **Plotting** etc. The first part requires the identification and handling of missing or null values. For example, if there are videos with missing categories or titles, decide whether to fill these gaps with default values or to remove these records, checking for and removing any duplicate records to ensure each entry in the dataset is unique, ensuring that each column is of the correct data type (e.g., dates should be in a date/time format, numeric values like views and likes should be integers or floats), looking for entries that don't make sense, such as negative numbers in views, likes, or comment counts, and handle them appropriately.

3.1 Description of Tasks

Here the possible uses of the dataset are discussed through the various visualisation techniques and analysis:

1. **Geographical trends worldwide**: The data organized in regional files help study video trends over that region.
2. **Sentiment analysis**: The video tags can be used to analyze the content and genre of the videos and their prevalence across different regions, generate word clouds and predict trends.
3. **Performance Metrics Analysis**: Determine what factors contribute to video popularity. Analyse the correlation between engagement metrics and popularity, identifying key predictors of success.
4. **Graphical Visualisation**: Analyse the spread of categories in a region, distribution of trending videos by categories, dislike and like count, etc.
5. **Global Visualisation**: Observe the spread of videos by region and the global Sankey diagram of the most trending videos from region to region.
6. **Temporal trend analysis**: For all the possible use cases discussed before, a time based statistical analysis is desired to know how, for example, the various metrics change over

months. This will be done by identifying statistical patterns and fluctuation over time.

3.2 Dashboard creation and software

1. We plan to use Plotly and Dash to develop interactive web dashboards that allow users to explore the dataset. Include filters for categories, periods, and regions to customize views.

2. For generating some of the static,animated, and interactive visualizations in Python, Matplotlib & Seaborn might be used.

3. Integrate visualizations for each analysis task, ensuring the layout is intuitive and highlights key insights.

4. Dash applications can be deployed to the web, making the dashboard accessible to a broad audience without commercial software licenses.

3.3 Preliminary experiments with dataset

All of the country-wise CSV files present in the dataset were merged and a category analysis was done to gain insights into the frequency distribution of all the unique categories present in the YouTube videos. It was observed that the only unique category IDs present in dataset were `10, 23, 24, 25, 22, 26, 1, 28, 20, 17, 29, 15, 19, 2, 27, 43, 30`. The bar chart (2) illustrating the distribution is given below:

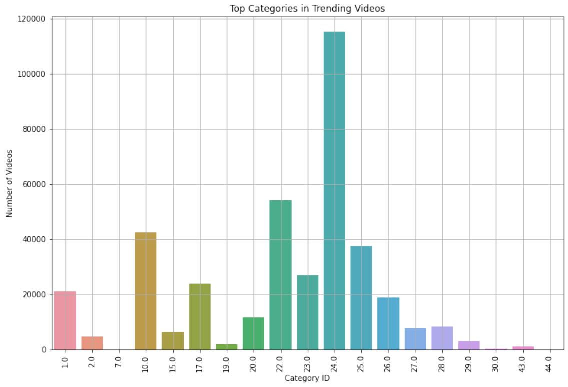

Figure 2: Frequency Distribution of categories

Next, the heatmap(3) shows the correlation between different numerical columns: `views`, `likes`, `dislikes`, and `comment counts`. It provides insights into how these metrics are related to each other. For instance, a high correlation between likes and views would indicate that videos with more views tend to also have more likes.

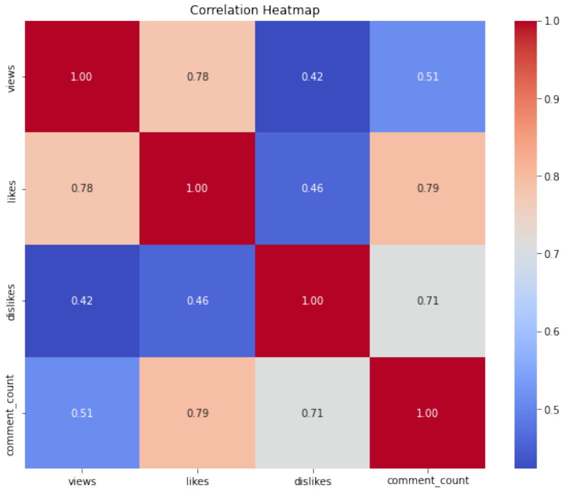

Figure 3: Correlation Heatmap of categories

The bar chart(4) illustrates the average views that videos in each category receive. Some categories may naturally attract more views due to their content's nature or popularity. This plot helps in identifying which categories are more likely to garner a higher viewership on average.

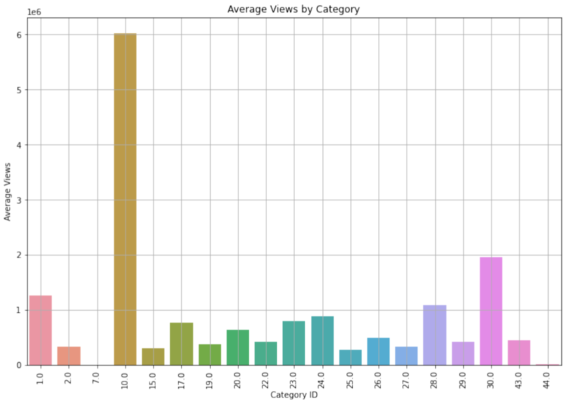

Figure 4: Average views per category ID

Following this we plan to clean the dataset, extract relevant features, perform the various tasks as mentioned.

4 Conclusion

To summarise, we have proposed an outline for exploring the YouTube trending video datasets. YouTube being an information hub, is used for all variety of reasons by people: consumers, creators and advertisers. Hence, it is critical to derive insights from its analysis, as it would help the various stakeholders make informed decisions and strategies to suit their goal. The structure of the dataset obtained from Kaggle was discussed. Application domain and use cases were elaborated with the various methodologies involved. The preliminary inferences from the data were discussed. Through structured data analysis of YouTube trending videos using open-source tools, valuable insights can be gleaned about viewer preferences, engagement patterns, and content trends. Use of ML for sentiment analysis and gaining insights would be challenging and exact methodology is yet to be understood. If required, necessary amendments would be made to the proposal as more insights are gained from the analysis in order to make our project more impactful.

5 Division of Work

The following table (1) denotes the individual task assignments for the creation of a visual analytics dashboard on the above mentioned dataset.

Tasks	**Member Responsible**
Sentiment Analysis	Armeet, Rajarshi, Ritam
Geographical trends, Graphical Visualisation	Ojsi, Sushmita
Global Visualisation	Armeet, Rajarshi
Performance Metrics Analysis	Sourit Saha, Ishaan
Temporal Trend analysis	Jatin, Yash
Dashboard Frontend	Ritam, Sushmita
Interpret findings, prepare report and presentation	Jatin, Yash, Ojsi

Table 1: Task Assignment

References

[1] Global YouTube Statistics 2023, https://www.kaggle.com/datasets/nelgiriyewithana/global-youtube-statistics-2023/data.

[2] Datasnaek YouTube New, https://www.kaggle.com/datasets/datasnaek/youtube-new/data.