

CSP 571 - Data Preparation and Analysis

Project - Proposal & Outline

“Exploratory Data Analysis on Trending YouTube Videos”

Group members

Venugopal Katragadda(A20544776)

Sannihitha Gudimalla (A20560248)

Sruthi Kondapalli (A20554780)

Bezawada Sai Sravanya (A20561552)

Project Proposal

1.1 Description of the Project and Research Goal

YouTube has become one of the biggest platforms for video sharing and consumption with billions of users worldwide. Everyday millions of videos are uploaded across various categories and only a small fraction of them make it to the **Trending** section. Understanding the dynamics of trending videos on YouTube can be super valuable for content creators, marketers and researchers to optimize video reach and engagement.

The goal of this project is to analyze the factors that make trending videos on YouTube popular. By doing an **Exploratory Data Analysis (EDA)** on historical trending videos and using modern tools, this project will uncover meaningful patterns and correlations in video success.

This project will look into various attributes such as **video categories, publishing time, user engagement metrics (likes, views, comments), sentiment analysis of video titles and descriptions, and clustering** to find trends. And by comparing past and present trends this project will also see if the factors that made a video go viral **five years ago** still apply today.

Through this analysis this project will develop predictions that can guide content creators and digital marketing strategists to optimize their video content for more reach and engagement on YouTube.

1.2 Research Questions

This project will explore the following key questions:

1. **Which video categories are most likely to trend on YouTube?**
2. **What is the optimal duration for a video to trend in different countries?**
3. **Does the time of day or day of the week a video is published affect its chances of trending?**

4. **Are there specific keywords or tags that increase the likelihood of a video trending?**
 5. **Can sentiment analysis predict whether a video is likely to trend based on sentiments expressed in the title or comments?**
 6. **Can clustering techniques help identify niche content categories that have a higher probability of producing trending videos?**
 7. **Are the same trends from the past still influencing trending videos today?**
-

1.3 Proposed Methodology and Approach

This project will follow a structured data-driven approach to examine YouTube's trending videos, applying data preprocessing, exploratory analysis, and predictive modeling. The methodology will include the following steps:

1.3.1 Data Collection

- Import data using **R** from publicly available datasets.
- Analyze a **five-year-old dataset** consisting of **10 CSV files**, each representing a different country, containing **16 columns** with relevant video details.
- Use the **YouTube API** or web scraping techniques to gather recent trending video data and compare past and present trends.

1.3.2 Data Preprocessing

- Clean the dataset by **removing irrelevant columns** and handling missing values.
- Perform **text preprocessing** on titles, descriptions, and comments for **sentiment analysis**, including removing stopwords, stemming, or lemmatization.

1.3.3 Exploratory Data Analysis (EDA)

EDA will help in identifying patterns and relationships in the data. The analysis will be conducted in the following areas:

1. Country and Category Analysis

- Explore trends across different **countries**.
- Analyze the relationship between **language usage** and trending video performance.
- Identify which **categories** are most likely to trend in each country.

2. Sentiment Analysis

- Clean textual variables such as **titles, descriptions, and comments**.
- Analyze the **distribution of sentiment** (positive, negative, neutral).
- Determine if **sentiment scores correlate** with video popularity and engagement.

3. Clustering Analysis

- Engineer new features to better understand video attributes.
- Apply **clustering algorithms** to group similar videos based on characteristics.
- Visualize the **clusters** to identify distinct content patterns.

1.4 Metrics for Measuring Analysis Results

This project will use various performance metrics to quantify the results of the analysis:

1. Exploratory Data Analysis

- Data **distribution**, measures of **central tendency**, and **variability** will be examined.

2. Sentiment Analysis

- The project will measure sentiment distribution (positive, negative, neutral) and calculate **average sentiment scores**.

3. Clustering Analysis

- The project will evaluate the quality of clusters using metrics like **silhouette score** and **inertia** to determine how effectively videos are grouped based on common characteristics.

2. Project Outline

2.1 Literature Review and Related Work

2.1.1 Data Repository

1. [YouTube Trending Statistics EDA \(Kaggle\)](#)
2. Scraped dataset from **YouTube API** for recent trending videos.

2.1.2 Reference Resources

1. [YouTube Trending Videos: Boosting Machine Learning Results Using EDA](#)
2. [YouTube Data Analysis & Prediction of Views and Categories](#)
3. [Trending Videos: Measurement and Analysis \(arXiv\)](#)

2.1.3 Supplemental Resources

- [Analyzing Popular YouTubers: An EDA Project Using YouTube Video Data](#)

2.2 Data Sources and Feature Descriptions

- **Dataset Type:** Multivariate
- **Dataset Size:** 10 datasets categorized by country, each containing **16 columns** and approximately **20,000 records**.
- **Countries Analyzed:** Canada, USA, Great Britain, France, Denmark, Russia, Mexico, South Korea, India, Japan.

- **Missing Values:** Yes

Feature Descriptions

- **video_id:** Unique identifier for each video.
 - **title:** Title of the video.
 - **trending_date:** Date the video appeared in trending.
 - **channel_title:** Name of the YouTube channel.
 - **category_id:** Category to which the video belongs.
 - **published_time:** Date and time of publication.
 - **tags:** Keywords describing the video.
 - **views, likes, dislikes, comment_count:** Engagement metrics.
 - **thumbnail_link:** URL to video thumbnail.
 - **comments_disabled, ratings_disabled:** Boolean values indicating whether comments or ratings are disabled.
 - **video_error_or_removed:** Whether the video has been removed from YouTube.
 - **description:** Video description.
-

2.3 Data Processing Pipeline

1. Cleaning the Data

- Handle missing values using **mean or median imputation**.
- Convert categorical **category_id** into **dummy variables**.
- Normalize numerical features (**views, likes, comments**) for consistency.

2. Data Transformation

- Convert **trending_date** and **published_time** into datetime format.
- Extract new features such as **day of the week** and **hour of publication**.

3. Outlier Detection

- Use **Z-score** and **IQR methods** to detect extreme values.

4. Sentiment Analysis & Clustering

- Perform **sentiment analysis** on titles, descriptions, and comments.
- Implement clustering techniques to **identify patterns in trending videos**.

2.4 Model Selection

- Identify **key features** that impact video popularity.
- Use **tree-based models** and **correlation analysis** to assess feature importance.
- Implement clustering algorithms like **K-Means** or **DBSCAN** to group similar videos.

2.5 Software and Tools

- **Libraries:** tibble, DT, knitr, tm, ggplot2, wordcloud, dplyr, fitdistrplus, plotly, plyr, textblob, cluster, YouTube API.
- **Software & Languages:** RStudio, Jupyter Notebook, R, Python (for YouTube API).

This project aims to provide actionable insights into YouTube video trends, helping content creators optimize their strategies for better audience engagement.