### Decoding YouTube Popularity: Predicting and Analyzing of Trending or Non Trending

### Team 2A - Ruicheng Li, Chun-Hui Tsai, Lalitha Bhupalam, Hilary Nguyen, Abhinav Samavedula

1. Introduction
   1. provides an overview of the project and justification for your choice of this topic. What questions that you try to answer in this project?

Youtube has become part of our daily life. From learning how to do organic chemistry, to travel vlogs to music videos, youtube has it all. It has become a platform that connects people. Therefore, it is no wonder that Youtube is the top video sharing and content creation platform with over 2.51 billion monthly users. Since it has so many users, it can be a little difficult to achieve popularity on youtube. Hence we decided to analyze trending and non trending videos from youtube.

These were are 4 main questions throughout this project, all of them stemming from our main objective of analyzing which factors influence the popularity of a video on Youtube.

What category do the trending videos on YouTube platform usually belong to?

What’s the best time in a day or week to upload videos?

What is the relationship between different metrics in a video, like views, likes, and comment counts?

Are there any key words or phrases that are common among popular videos?

1. (Related Work)
   1. anything that inspires you that may be some new article, a paper, a web site, a case study done by others or something that we discussed in class.

The inspirations for our topic:

1. According to an article by Connections Academy in 2022, when children between the ages of 8 and 12 are asked what they want to be when they grow up, many are saying “Youtuber” more now than ever. This leads to question…
2. So, how does one go viral to make a living off of Youtube? A case study from 2020 called “Data-Driven Tips to Make a Video Go Viral on YouTube” studied data in the US market and which channels are the most popular, most viewed videos, most popular categories, etc. This inspired us to study up-to-date Youtube data of our own to answer some questions for exploration.

<https://medium.com/swlh/data-driven-tips-to-make-avideo-go-viral-on-youtube-e7117f51fc2f>

<https://www.connectionsacademy.com/support/resources/article/career-goals-among-children/>

1. Data
   1. including details on how you get the data such as the source of the data, the scraping method and so on. You may also want to detail how you clean up the data.

For the Data source, we have obtained the trending data from Kaggle Dataset which is updated everyday. It uses Youtube Data API version 3 to collect data by using the chart parameter set to most popular. This collects all the data of trending videos for each region on that day. The dataset we found has daily data from 2020-2023 present day.

For the non -trending data, we did not have a particular dataset. So we decided to web crawl a website called DevPicker.com. This website has a feature that generates random videos.

Since this feature picks very unseen videos that most people haven’t come across, we thought it would represent the non-trending data well. So we replicated the Http requests of this website to get the random video Ids and then used Youtube Data API to get the rest of the details about the video. Finally we checked if the video was published in the years 2020 - 2023 to ensure that the non-trending data matches the timeframe of publish as the data from the kaggle dataset.

Once we have both trending and non-trending data we combined them into a single dataframe adding a label field that captures whether the video is trending or non-trending . Then we shuffled the order of the rows to create our final dataframe.

Coming to the data descriptions we have 3865 rows and 16 columns. Each row contains all the details of a video such as video and channel ids, titles, descriptions, engagement statistics etc. Also our class distribution is almost equal to 50 percent of each class.

video\_id

channel\_id

channel\_title

category\_id

trending\_date

publishedAt

tags

view\_count

likes

dislikes

comment\_count

thumbnail\_link

comments\_disabled

ratings\_disabled

description

label (trending:1 /non-trending:0)

Data cleaning

Removing duplicates because web crawling the random youtube generator too many times means that we get some duplicates

Removing channel id, thumbnail link and 'video\_id','channelId','thumbnail\_link','comments\_disabled','ratings\_disabled'

Cleaning tags to remove the vertical bar - |

Making all the dates: published date and trending date be in the same format y m d

Converting category id to actual category names

1. Exploratory analysis
   1. A description of the statistical or machine learning methods that you use in this project, using the content taught in class but you are not restricted to coursework.
2. Data visualization
3. Classification

Tried two models

Data preprocessing

Model 1

Model 2

Feature importance

1. Text analysis
2. Analyses of the data
   1. using the methods discussed above (again, you are free to bring in any relevant methods, whether covered in class or not).

Data visualisation

Classification

Text analysis

1. Conclusion
   1. A brief concluding section that highlights the main elements of your analysis and findings. What can we learn from your project? What answers did you find for the questions that you proposed?
2. Summary of findings

Learnings

Answers for the four questions

Alex, Chun Hui, Lalitha, Hilary, Abhi - Coding

Lalitha, Abhi, Hilary - Report

Sunday, Nov 12th - Coding

Thursday, Nov 16th - First Draft

Thursday, Nov 23rd - Second Draft + Presentation

Sunday, Nov 26th - Final Draft

Youtube api key: AIzaSyC-M-oJhyb2iJJu3gQsl6bje3VnpJAsRRU

[YouTube Trending Video Dataset (updated daily) (kaggle.com)](https://www.kaggle.com/datasets/rsrishav/youtube-trending-video-dataset) new dataset