US YouTube Videos Trend analysis & like prediction

# **Vikas Ranjan DSC680-Applied Data Science ranjaninv.github.io**

# Which Domain?

This project is intended to be focused on social media domain. In my pursuit to find the topic, I was intrigued by online video sharing and social media platform YouTube. YouTube is the world’s second largest search engine and second most visited site after Google. In fact, 37% of all mobile internet traffic belongs to YouTube. It is the second most popular social media platform with almost 1.9bn users. More than 500 hours of video are uploaded to YouTube every minute. We watch over 1 billion hours of YouTube videos a day, more than Netflix and Facebook video combined. It has not only transformed the music industry, but on a broader level it has given power to its views. It has not only become a great alternative to traditional media. Content creators are being paid to create and upload videos which is based on the number of subscribers they have.

To determine the year’s top-trending videos, YouTube uses a combination of factors including measuring users interactions such as number of views, shares, comments and likes. It is remarkable to note that they are not the most-viewed videos overall for the year. Top performers on the YouTube trending list are music videos, celebrity and/or reality TV performances.

All these facts propelled me to look for and find the dataset which has statistics on YouTube views. The intent of this project is to perform EDA and statistical analysis on the dataset. I’ll also look at various factors that would affect on how popular a YouTube video will be. I’ll also be applying machine learning algorithm to predict if a video will get more than 1M views or not. I’ll also perform sentiment analysis on the comments of these videos.

# Which Data?

What is the dataset you’ll be examining? Please provide a codebook if there is one or a link to the dataset as well as a detailed description.

I extracted the dataset from Kaggle from below url:

<https://www.kaggle.com/datasnaek/youtube-new>

This dataset includes 40950 rows and 16 feature variables. Each row corresponds to a unique video, and includes the variables:

Video\_id: Unique video ID.

Trending\_date: It is the trending date.

Title: Title of the video.

Channel\_title: Name of the channel under which video was uploaded.

Category\_id: Category of the video.

Publish\_time: Date on which the video was published.

Tags: Free form tag attached to the video.

Views: Number of views associated with the video.

Likes: Number of likes the video has.

Dislikes: Number of dislikes the video has.

Comment\_count: Number of comments on the video.

Thumbnail\_link: Link of the thumbnail.

Comments\_disabled: Boolean field to suggest if comments are disabled on the video.

Ratings\_disabled: Boolean field to suggest if Ratings are disabled.

Video\_error\_or\_removed: Boolean field for the error with the video.

Description: Description of the video.

# Research Questions? Benefits? Why analyze these data?

There are many benefits of data analysis. It does help with finding hidden trends and patterns in the data. These insights help with understanding which categories and channels gain most likes and what are the keywords which identify with trending videos. Graph analysis helps that significantly. That would also tell which attributes are corelated. They would also help to guide future business decisions.

Which categories have most liked videos the most?

Which channels have most likes and dislikes?

How positive or negative comments are on top channels?

Which are most view channels and what’s the average view on the videos of those channels?

# What Method?

I’ll start with data cleanup. Once data cleanup is completed, I’ll perform EDA (Exploratory Data Analysis), and try to find the trends and patterns in the data based on the categories, % of like and dislikes. I’ll use visualizations and graph analysis to do the same. I’ll also build word cloud on the description, comments, and tags. I’ll also perform sentiment analysis on tags, comments, and description.

I’ll use linear regression classification model on likes field and validate its accuracy based on the r-square and mean absolute error.

# Potential Issues?

Data quality may come across as a potential issue. If I see too much of missing data and data integrity issues, I may have to look for another dataset or spend too much time in data cleanup which may delay my project. I also don’t have the dataset which all the parameter which would play an important role in determining if a video will get enough likes.

# Concluding Remarks

All in all, this project will allow me to apply the skills I learnt during this course. I’ll be performing data cleanup, EDA and graph analysis to gain insight from the data. I’ll be performing sentiment analysis on the comments fields to understand subscriber behavior and reception to the views. And will be applying machine learning to predict if a video will get enough likes or not. This will help to guide business decision for future from the YouTube perspective and from the content creator/influencer perspective.

# References:

<https://www.kaggle.com/datasnaek/youtube-new>

<https://towardsdatascience.com/youtube-views-predictor-9ec573090acb>

<https://medium.com/analytics-vidhya/youtube-view-prediction-with-machine-learning-fdd4f40f352d>

<https://arno.uvt.nl/show.cgi?fid=137707>

<https://blog.hootsuite.com/youtube-stats-marketers/>

<https://popsters.com/blog/post/youtube-statistics-for-any-channel>

Applied Text Analysis with Python, Benjamin Bengfort, Rebecca Bilbro & Tony Ojeda

Machine Learning with Python Cookbook, Chris Albon

<https://www.digitaltrends.com/features/how-youtube-has-changed-the-world-in-15-years/>

<https://www.brandwatch.com/blog/youtube-stats/>