Heinz 95-845: Final Project Proposal

Nettie Lian JIAOYINL/jiaoyinl@andrew.cmu.edu
Heinz College of Information Systems and Public Policy
Carnegie Mellon University
Pittsburgh, PA, United States

Stephen Lin DONGLIEL/dongliel@address.edu
Heinz College of Information Systems and Public Policy
Carnegie Mellon University
Pittsburgh, PA, United States

1. Project Description

YouTube is the world's most famous video-sharing platform.The average number of YouTube video views per day is about 2 billion. Only 0.33% of them have over 1 million views while 53% of YouTube videos have fewer than 500 views. [1] Our study attempts to provide an overall characterization of YouTube by:

- Identifying factors affect a video's potential to acquire viewership.
- Develop and test models to predict a new video's viewership based on the channel's previous performances and identified factors.

Likely outcome of the analysis, ceteris paribus:

- Video titles containing popular words will likely attain popularity, turning into high views.
- Videos uploaded by popular channels will likely attain popularity, realizing high views.
- Videos in games, sports, comedy categories will like attain popularity.

Most prediction models on YouTube video viewership involve a spectrum of feature engineering procedures, including image processing on thumbnails, text-mining on titles, tags, descriptions on top of ordinary feature engineering process. We contend that the pre-processing procedure aforementioned is overly-complicated, requiring resources, technical talent, and time beyond control of the general public and small businesses. Our analysis, on the other hand, draws on feature selection insights from *Bartl, 2018*[7], aims to provide equivalent predictive power with simpler model designs and greater interpretability.

Most prior studies have focused on predicting video popularity by developing an elaborate model encompassing a spectrum of predictors: including thumbnails [2][3], dynamic view count trajectories [2][3][4][5], and other video parameters. Our study will emphasize on using static features, complemented by additional features extracted by text mining techniques on video titles, and aim to develop a convenient, easy-to-replicate predictive model. *Bartl, 2018* have conducted analysis on determining factors that will affect viewership. Our

study, building the simplest model to predict viewership, will be able to validate on finding of *Bartl, 2018*, and benchmarked against *Mekouar et al., 2017* [8] and another study[9].

Given that retrieving and analysing data on all and most up to date videos posted on YouTube is impossible, this study will use a dataset which contains several months of data on daily trending YouTube videos. The dataset was found on data sharing website Kaggle.com and was most recently updated in June 2019. Each record represents a different video that was uploaded to YouTube. The outcome Y will be views count prediction, V covariates are title, channel_title, category_id(describes the category of this particular video in numeric format), publish_time(time which the video was published in date format), tags and description(words or phrases that can be included in the video's description box, which let the viewers know what the videos are about and help the videos rank higher in search results), likes and dislikes, comment_count(the number of comments for this video), thumbnail_link(the thumbnail for this video in picture format), comments_disabled and rating_disabled(whether this youtuber permits viewers to leave comments or submit likesdislikes for this video, in boolean), video_error_or_removed(whether this video has been removed or disabled in boolean format).

It's a non-randomised study because the dataset only contains daily top trending videos instead of randomly chosen videos from all uploaded videos. Dataset includes the video title, tags, and description, which are all represented in a context text form. In order to understand and normalize our data, we will use text mining techniques to generate the desired information from context. There are also some missingness in the datasets which requires MICE to take place. Feature scaling is also required for this dataset, which limits the range of features so they are comparable. Upon the completion of preprocessing of data, we will then perform a set of techniques to do feature engineering and feature selection. Stepwise, forward, and backward techniques are all applicable for our model. Simple linear regression, polynomial regression, support vector regression, decision tree regression, and random forest regression can all be used for the purpose of this project. The dataset has about 80,000 rows and 15 features, which is an appropriate size for a course project. Since the study is trying to make predictions of the number counts of views, AUC, ROC, and MSE are all applicable to measure the performance of the model. For the purpose of this study we will use MSE to measure the goodness of the trained model.

Many of existing literatures have explored the use of image processing techniques to extract information for thumbnails. (a, b) indicates that thumbnails can be positive correlated to video popularity. Due to expected timeframe of the project, we will focus on extracting information for titles using text mining techniques. Furthermore, prior studies have utilized dynamic statistics, tracking the growth of viewership over time to predict popularity. We do not have access to time series data based on the way data was collected. Another major limitation of the analysis is the lack of external linkages. Videos can go viral when it is related to news. While text-mining on titles and description can account for buzzwords in the past, we lack means to identify rising buzzwords.

We have envisioned that the pipeline can be utilized by organizations who wish to place advertisements, or to sponsor videos based on popularity. The analysis will be particularly useful to decide which videos to place ads on because it requires only static parameters collected at the time videos are uploaded. It can be indicative for sponsorship opportuni-

ties as the predictive model will be able to predict viewership of next video based on the popularity of the channel.

References

[1] *Hecht, Ed. "Only 0.33% of YouTube Videos Generate 1 Million or More Views... SXSW." The Trichordist, 12 Mar. 2014, thetrichordist.com/2014/03/12/only-0-33-of-youtube-videos-generate-more-than-1m-views/.*

[2] *Giulia Fontanini, Marco Bertini, and Alberto Del Bimbo. 2016. Web Video Popularity Prediction using Sentiment and Content Visual Features. In Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval (ICMR '16). ACM, New York, NY, USA, 289-292. DOI: https://doi.org/10.1145/2911996.2912053*

[3] *Jiang, L., Miao, Y., Yang, Y., Lan, Z., Hauptmann, A.G.: Viral video style: a closer look at viral videos on YouTube. In: Proceedings of ACM International Conference on Multimedia Retrieval, p. 193 (2014)*

[4] *Figueiredo, Flavio & Benevenuto, Fabrício Almeida, Jussara. (2011). The tube over time: Characterizing popularity growth of YouTube videos. Proceedings of the 4th ACM International Conference on Web Search and Data Mining, WSDM 2011. 745-754. 10.1145/1935826.1935925.*

[5] *Choe M.G., Park J.H., Seo D.W. (2019) How Long Will Your Videos Remain Popular? Empirical Study of the Impact of Video Features on YouTube Trending Using Deep Learning Methodologies. In: Xu J., Zhu B., Liu X., Shaw M., Zhang H., Fan M. (eds) The Ecosystem of e-Business: Technologies, Stakeholders, and Connections. WEB 2018. Lecture Notes in Business Information Processing, vol 357. Springer, Cham*

[6] *Blundell, R.,Griffith, R., &Van Reenen, J. (1995). Dynamic Count Data Models of Technological Innovation. The Economic Journal,105(429), 333-344. doi:10.2307/2235494*

[7] Bärtl, M. (2018). YouTube channels, uploads and views: A statistical analysis of the past 10 years. Convergence, 24(1), 16–32. https://doi.org/10.1177/1354856517736979

[8] *Mekouar, Soufiana & Zrira, Nabila & Bouyakhf, El Houssine. (2017). Popularity Prediction of Videos in YouTube as Case Study: A Regression Analysis Study. 1-6. 10.1145/3090354.3090406.*

[9] *Anonymous Author(s). https://pdfs.semanticscholar.org/7dad/e77c5a6c58ec2543ea10ed499395957fbcf4.pd*