# report

October 1, 2019

Project Summary: •I want to analyze trends of youtube's trending videos to determine some of the relating factors that lead to a YouTube video being added to trending. Relatedly I want to analyze the categories that seem to become trending the most as well, potentially also looking at whether "clickbait" titles are more likely to reach the trending page of YouTube. To do this I will use a combination of views and the like to dislike ratio to try and determine a "success_rating". My target variable will therefor be this success_rating whose function is still being calculated. However It will probably be a weighted sum of views and likes and dislikes. •This is mostly a classification problem. However I could also take an unsupervised approach and see if the categories I selected as impactful are actually what are driving these videos to trending. •I think this is very interesting since most of the media I consume is from YouTube. YouTube is also becoming a space for budding talent to gain a lot of exposure, so seeing what factors lead people to gain popularity is a relevant conversation. Especially as more and more people can start using YouTube as full time careers due to revenue from advertisements and sponsorships and patreons. (a money donating app that lets you pledge to creators giving them a monthly salary of around 1 to 2 dollars per person up to 100s). Kaggle has some suggested uses for the data I seem to be doing a sort of combination of things and looking into what makes a trending video a trending video: " Possible uses for this dataset could include:

Sentiment analysis in a variety of forms eg using language analysis to determine most probable reaction to trending videos Categorising YouTube videos based on their comments and statistics. This couild be interesting to see if there were certain categories that caused more vigorious discourse and reacitons Training ML algorithms like RNNs to generate their own YouTube comments. This is self explanitory Statistical analysis over time. One of the possible columns is dates so it is feasable to look at how categories and opinions have changed over time. " Dataset: •I was thinking of just using the dataset from the united states, however I have access to data sets from up to 8 other countries, so it could be an interesting project comparing the results among them to see if there is a significant difference between the type of trending videos that appear in other places. The US data set is 40000 videos long with 16 different columns. •The categories are pretty self-explanatory they are as follows: o video_id : this is a unique string meant to identify the video, this can be used a key since it is unique o trending_date: this is a string of the date the video was first put up on the trending page o title: This is a string of the name of the video o channel_title: This is a string of the name of the channel that posted the video. o category_id: This is a list of keys that relate to a .json file that has the connected categories. o publish_time: This is the date that the video was published, this can be very different from the time it went trending which can be very interesting o tags: This is a string of the tags associated with the video. o Views: Integer value of the number of views o Likes: Number of likes o Dislikes: Number of dislikes o comment_count: Number of comments o thumbnail_link: link to the video o comments_disabled: Boolean whether comments were allowed for the video o ratings_disabled: Boolean on whether ratings where ddisplayed o

video_error_or_removed: Boolean whether the video was removed o description: The description provided by the publisher underneath the YouTube video. •The data set is from Kaggle and some of the suggested uses for the data are: o Sentiment analysis in a variety of forms o Categorising YouTube videos based on their comments and statistics. o Training ML algorithms like RNNs to generate their own YouTube comments. • Of all of the things suggested sentiment analysis sounds very cool however I do not have comments or know how to successfully implement sentiment analysis, however it may be similar to what I plan on doing with movie titles such as determing the clickbaityness of video titles. Preprocessing: A lot of the data is very usable. I used StandardScaler on views Likes and Dislikes because there was no clear Min or Max to use here, I felt it would benefit more from not having a ceiling and all that really matters is how they are scaled with each other, which is perfect for standard scaler. I used LabelEncoder on trending date since a lot of the dates are the same and it will make it easier to keep track of them. Currently I left title as is, however I am tempted to break title down even further into a boolean by deciding whether it is considered a click bait title or not (there is a prebuild algorithm for this). I used label encoder on whether comments were disabled since this will allow me to get a nice 0 or 1 for true and false.

https://github.com/wward97/data1030project

[ ]: