Youtube daily trends and the algorithm that decides them

Charles Kash

The motivation for my project is for data analysis into the construction of the algorithms that decide whether a YouTube channel is likely to get their video in trending. It is no secret that the trending page is geared to show off various channels more often than others because of their own personal biases, but I wanted to get a better idea of how the various statistics that can be taken from the particular trending videos of a channel within a certain period of time, and then compare it with data readily available online to better understand how other channels can have more trending videos. The importance of having a trending video in YouTube is that you are able to advertise your channel and grow your channel much faster, so YouTube will be preferential to channels that they want to see grow.

The problem is that I am trying to understand if YouTube has any biases towards particular corporations or creators. There might be a relation to the political leanings of the creators that decide whether they are more likely to have a video on the trending page. I want to compare the trending data with the views data, the number of comments on the videos, the likes and dislikes, and see the correlation that each stat has to decide whether a video is trending. It is important to understand these biases if you want your channel to grow, but it also has legal ramifications for YouTube for if they push their algorithm to far to specific content it can be taken as curating the content that is published on their site. Being considered a publisher rather than a platform would allow for direct action to be taken against YouTube for the content on their site.

There has been plenty of studying into the YouTube algorithm for the sake of people keeping their channels alive, and with various sponsors pulling their ads from YouTube videos it is becoming more and more necessary to conform you video to best be noticed by the algorithm, whether that means associating your channel to particular age groups, political groups, genre of content and format of the video. There was an ex-Google engineer that had a method of extracting data from the site and would simulate the behavior of a YouTube user. It was during the presidential debating season where there would be plenty of political trending videos. The program would test the search algorithm to see which videos were more likely to be shown on the front page of google and in the recommended search terms, and the data was divided into either “Trump” or “Clinton”. There was a finding that there were particular biases to mainstream news clips and comedy clips with clear political biases but supposedly were supposed to be “even-handed”.

The dataset used has the statistical data of every trending YouTube video within a two-month or so period. The data that I took from the dataset was what I thought would be most relevant for a video to be trending, which would be in the same way other social media platforms have their trending topics. I would categorize each channel by how many trending videos they had gotten. I would categorize channels in the same way with respect to the total views that their trending videos got. The same would be done for likes and dislikes on their videos and for the number of comments on a particular video. The data was augmented into a new merged data frame and correlation coefficients were obtained from how each of the data types affected the amount of trending videos a channel had. A new dataset was created using the correlation coefficients with randomly generated numbers for the statistics to have a neural network test the data. The pass/fail condition was for the channel to pass 0.66, as the data had to be limited to represent the limited number of trending videos within a certain amount of time.

The training set and the test set were divided using the standard test size of 0.2 of the total data. The data was scaled and standardized from sklearn to be Standard Scalar. Four our neural network the classifier used was sequential, the optimizer was adam, the loss was binary crossentropy, and the metrics were for accuracy. The neural network was trained with a batch size of 10 and had 100 iterations. 1.000 accuracy was finally obtained after 80 iterations of the neural network learning. The importance of finding this information with the random set of datapoints is that now when you can create a dataset from recent trending channels and other channels that have high view rates and like-to-dislike ratios, the efficiency of the neural network can be tested to see if it in fact takes more iterations to get to 1.000 accuracy. That would mean that there is a particular bias for the trending channels that is not based on the merit of the video.