REPORT

IDS 561

Final Project

Submitted by -
David Kern
Debdeep Ghosh
Karan Gujrati

Introduction

Youtube is the largest video sharing website in the world, with over 70,000 years of content uploaded and counting. Its viewership is over one billion hours per day, larger than Netflix and Facebook combined. Content creators who upload videos on the site naturally look for any edge to increase viewership and interaction to reach the trending list, with calls to 'like and subscribe' becoming ubiquitous on the site. This has also led to the adoption of 'clickbait' titles, which contain heavily capitalized words and withhold information to entice users to click on the video even if they do not intend to watch it in its entirety. In this project, we attempted to capture this practice through sentiment analysis of video titles compared to other text present in the video, and determine whether this practice is effective at driving views and engagement.

Data Description

We analyzed all trending videos in the US and Great Britain from the year 2019. The data was taken from a Kaggle non-competition set present in the public domain. The full dataset consisted of 500MB of CSV files which included several countries, but we chose to use the US and UK as the number of records was relatively similar between the two (roughly 40,000) and results were mainly in the same language (English), making comparative sentiment analysis much easier. The dataset included 16 columns, with both numeric (views, length, publish date) and text (title, description, channel title) data. Several of these columns, such as publishing date, thumbnail link, and video error, were unnecessary and were dropped while cleaning the dataset.

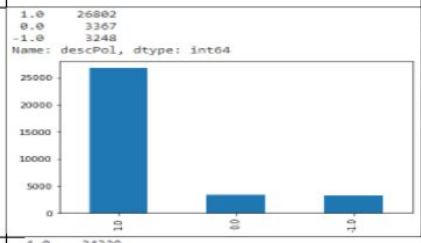
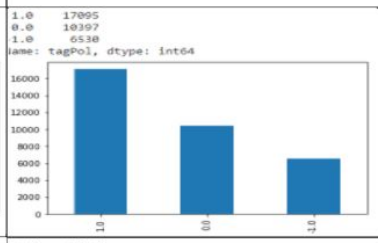
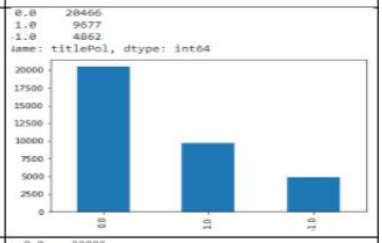
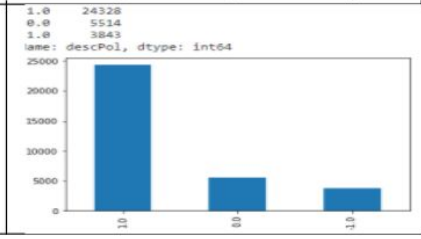
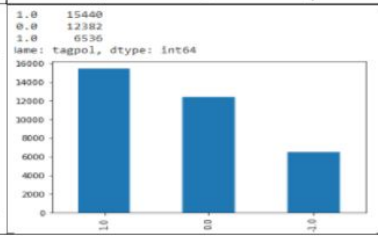
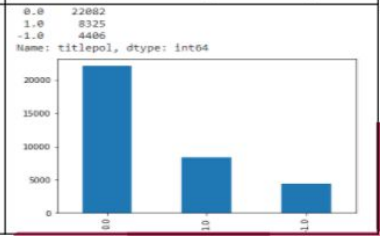
A link to the dataset can be found here: https://www.kaggle.com/datasnaek/youtube-new

Methods

First, we imported and cleaned the US and UK datasets using Spark. Some of the records had important fields such as views, likes, and dislikes with missing values which needed to be removed. Additionally, we needed to make a subset of the data when analyzing comment sentiment as many videos had comments disabled and we did not want to lose these records when examining other metrics. Finally, we prepared metrics to test sentiment scores against, as many of the numeric fields in the dataset were either irrelevant (length, date uploaded) or multicollinear to views (likes, dislikes, comments). We decided to transform these latter values into ratios that might have bearing on reaching the trending list (like/dislike ratio, like/view ratio, comment/view ratio).

We then implemented a sentiment analysis on the three main text fields available: tags, description, and title. While the category was also a text field included in the dataset, Youtube categories are chosen from a dropdown list and their wording is therefore out of the user's control when posting a video. Sentiment scores were calculated using the TextBlob package to slice and analyze words from the full-text fields. This analysis ignored capitalization and special characters, as the dataset did not have a large enough vocabulary to examine all caps vs normal capitalization in earnest. The sentiment score output was in the form of a polarity score, which is calculated by taking the count of positive sentiment words minus the count of negative sentiment words in a field and dividing by the total number of words. This gave a score between -1 (all words negative) and 1 (all words positive), which we could then use to perform regression against our metrics chosen earlier to find possible correlations.

Results

	Description	Tags	Title
US			
UK			

Sentiment distribution was relatively similar across the US and UK trending lists, with US sentiment remaining slightly more positive across all three text fields. As seen above, descriptions were overwhelmingly positive, tags generally positive, and titles neutral. This is likely tied to the purpose of each field; the description, the most verbose field, is intended to give extra context and reinforce an audience that has already been captured, but the title is meant to bring in a new audience with a few short words. This difference in sentiment shows us the trending videos might have a large amount of click-baiting in their titles, with a smaller proportion of positivity that would accompany this practice.

Sentimer	views	ldr	liker	commr
usdesc	0.523	0.632	0.874	0.581
ustag	0.066	0.261	0.981	0.791
ustitle	0.184	0.190	0.910	0.793
ukdesc	0.370	0.416	0.114	0.161
uktag	0.822	0.887	0.344	0.992
uktitle	0.368	0.919	0.528	0.353

The table above shows the p-values of correlation tests between the calculated sentiment scores and our metrics (views, like/dislike ratio, like/view ratio, comment/view ratio). All of them are greater than .05, showing that click-baiting might not drive viewership or engagement as much as conventional wisdom would suggest. The one caveat to this is the score between US tags and views is very close to .05, and further analysis across different dictionaries of sentiment could show a significant relationship. Interestingly, this relationship was inverse, meaning that negative tag sentiment could lead to higher viewership. Further analysis will be necessary to show whether this is unique to the 2019 dataset or a possible feature of US trending videos in general.

Conclusions/Further Study

Negative sentiment in video text fields does not appear to be a factor that drives viewership or engagement amongst trending videos in any large way. This does not necessarily mean that clickbait is entirely ineffective, however. Since all of these videos were taken from the trending list, it is impossible to tell whether this is a feature unique to the trending list, i.e. clickbait has diminishing returns on highly visible content. The second dataset of non-trending videos would allow analysis to generalize to all videos, rather than this reduced list. Additionally, a larger dataset of English language video text would allow analysis by capitalization, one of the most noticeable clickbait tactics.