YouTube Comment Sentiment Analysis & Video Growth Trends

Brenner Swenson – brenner.swenson@city.ac.uk - PRD1 A 2019/20

I. Analysis Domain, Questions, Plan

Globally we watch over one billion hours [1] of YouTube videos every day. There are over 400 hours [2] of video uploaded every minute, which equates to ~66 years of content each day. Each video on YouTube has a number of likes, and a number of dislikes. Together you can calculate a like to dislike ratio, which is the metric of interest for question 1 of this project. This ratio provides an indication of how the Internet feels about a specific video and is heavily used by self-proclaimed YouTubers to measure their fanbase's reactions to videos. Often YouTube comments are nice, but as is common on the Internet, comment sections can devolve in to discourse behind the guise of anonymity. Could data on the sentiment of the comments section be statistically related to the like/dislike ratio? Put more formally, the following question will provide the motivation for one half of this report:

1. To what extent is the median sentiment of YouTube video comments related to or a predictor of the like to dislike ratio?

Additionally, YouTube videos can become "viral" very quickly, meaning that a video's view count can reach millions, or even tens of millions of views in a single day. Alternatively, videos can grow slowly over time and still reach high view counts, or not grow at all. This leads to a second research question, which is the focus of the other half of the report:

2. What trends can we observe in how viral YouTube videos grow, and what characteristics contribute to a video's success (or lack thereof)?

To answer question 1, data needed to be collected via web scraping. Unfortunately, when a user visits a YouTube video, the comments are not rendered by default; a user only sees the comments after they scroll to the bottom of the page. This creates an issue for web-scraping Python libraries like BeautifulSoup [3] that are designed to parse the HTML only, and not interact with or manipulate the page in any way. To this end, an external GitHub repository [4] was leveraged to assist in dynamically extracting comments data from YouTube video pages. The plan was then to utilize this package, modify the code to this specific use case, and extract sufficient samples of comments data along with relevant video metadata from a random sample of approximately 3,300 YouTube videos.

To prepare to answer question 2, data collection was also needed, but a different kind of data. Comment data can be obtained at a single point in time, whereas to answer a question regarding the growth of something, a consistent sampling process over time is required. In contrasts to the comment data collection process, the video metadata such as views, likes, dislikes, channel name, date published, etc. are available via YouTube's API. A Python wrapper [5] for the API was employed to streamline the data collection process, which was planned for every hour of a two week period on a random selection of approximately 2,000 YouTube videos to observe how their characteristics develop over time.

II. Analytical Process

Both the comments dataset and the growth dataset were dependent on random samples of YouTube videos. The video URLs used to generate both datasets were scraped, using words sampled from the English language's 20k most popular words [6] to generate search queries, where the HTML of the search results was then parsed to extract the video URLs. Out of the parsed URLs, a random URL was then chosen; this process was then repeated n times depending on the quantity of videos to be extracted.

***Question 1 (sentiment)*:**

Upon completion of 14 web-scraping sessions for comments, the combined dataset contained 1.2 million unique comments across 3,297 YouTube videos. The distribution of comment languages prior to non-English language removal are displayed in Figure 1. Online comments are very messy, i.e. they are very inconsistent in length, their sentence structure highly varied, they're full of emoticons and slang,

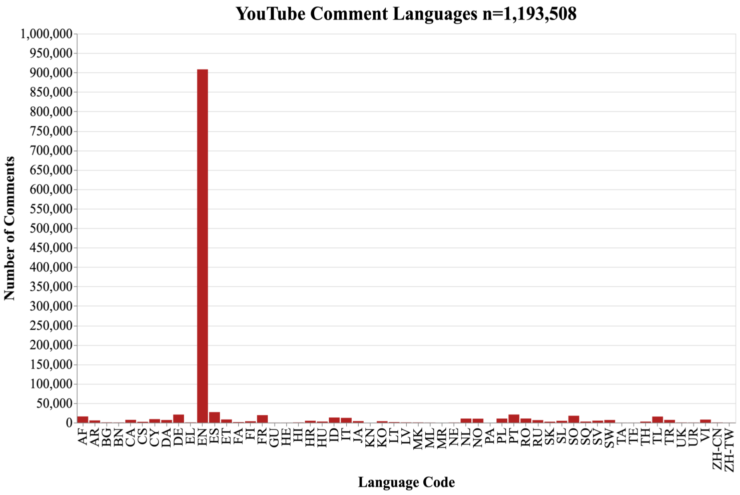


Figure 1. (right) Distribution of comments by their detected language. English represents 74.2% of all comments.

etc. For this reason, all comments were converted to their lemmas and lowercased, any emoticons and emojis converted to their respective English interpretations, and a small subset of stopwords removed.

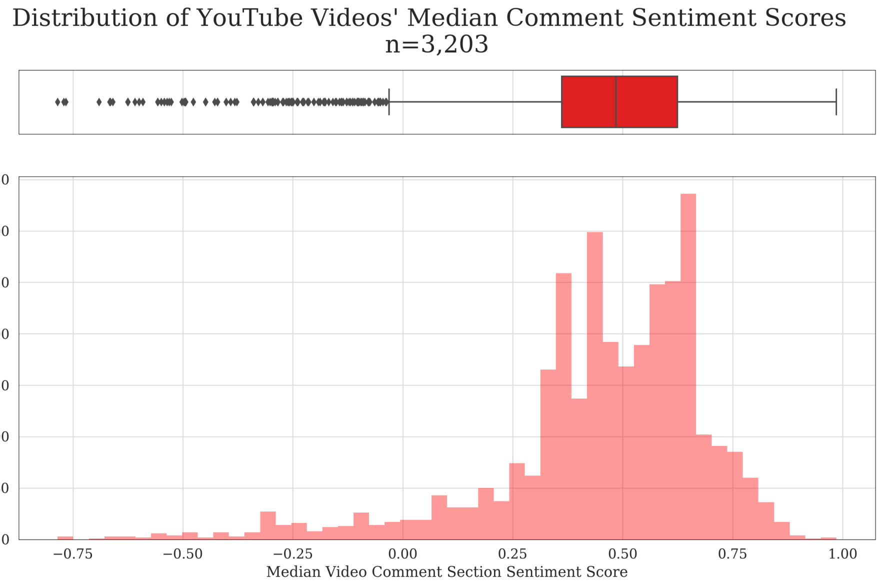


Figure 2. Median sentiment of YouTube comment sections by video.

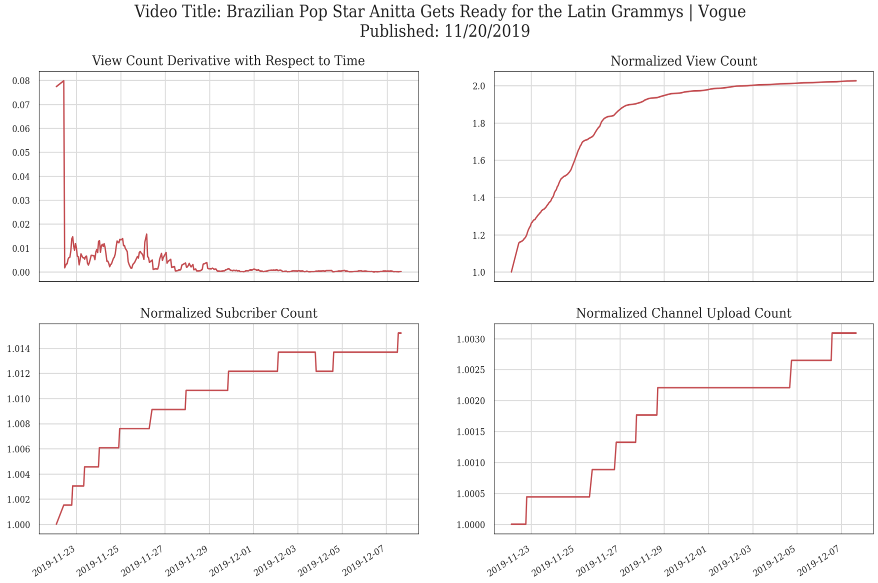


Figure 3. Selected features of a sampled viral video plotted over the sample period.

NLTK’s SentimentIntensityAnalyzer [7] was used to calculate the compound sentiment probabilities for each individual comment. As many comments are limited in length, or are generally ambiguous as to their underlying emotion, a large majority of the comments were labeled with 0 sentiment, meaning they are neither positive nor negative.

These comments were subsequently removed from the analysis. The median comment sentiment was then calculated for each video, which can be observed in Figure 2 to be positive for the large majority of videos.

Question 2 (growth):

One of the methods YouTube channels often use to gain as many views as possible is to create what are known as "clickbait" titles; i.e. a title that makes a potential viewer more likely to click the video. Clickbait titles are often all uppercase or make outrageous claims.

To attempt to measure this phenomenon, **pct_title_uppercase** was created, a feature that conveys the number of uppercase characters as a percentage of all alphabetic characters in the video title.

Additionally, the gradient of the normalized viewcount **view_growth** feature was taken with respect to time. As the video statistics were collected at non-standardized time intervals, the data were required to be resampled and interpolated to create an even index for gradient calculation. NumPy's gradient [8] function was used to calculate the derivative by central differences in turn creating the **viewcount_slope** feature, which can be seen in the top-left graph in Figure 3 for a sampled viral video.

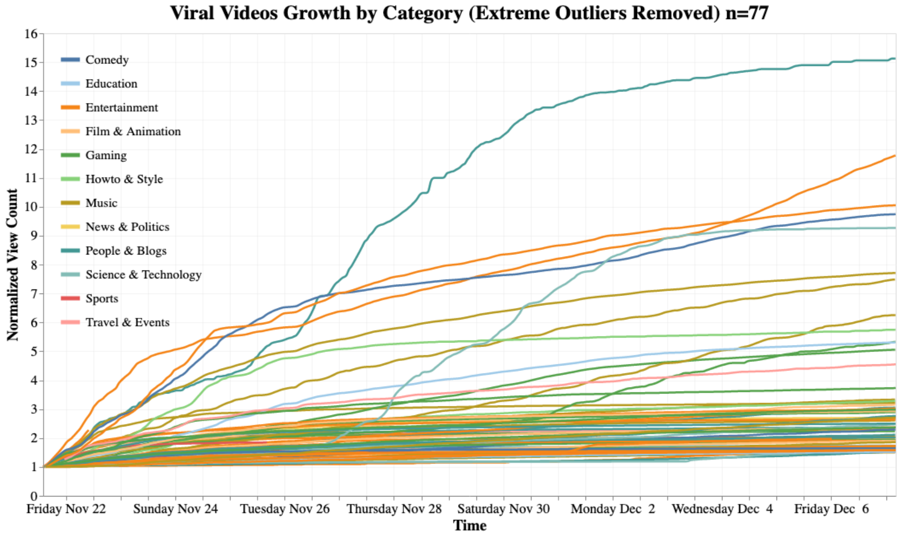


Figure 4. Temporal evolution of viral videos' normalized view counts by category.

Many more features were engineered and are found in the accompanying computational notebook. As the research question is interested in viral videos, the randomly sampled dataset needed to be reduced to only analyze the videos that had grown significantly over the sample period (1.5x), were published in within a recent time period (60 days for this study), and amassed more than a threshold (50,000) total views by the end of the sampling period.

Figure 4 displays the view growth of the selected viral videos over time broken out by category, where music and entertainment videos are seen to comprise the highest portion of viral videos.

III. Findings and Reflections

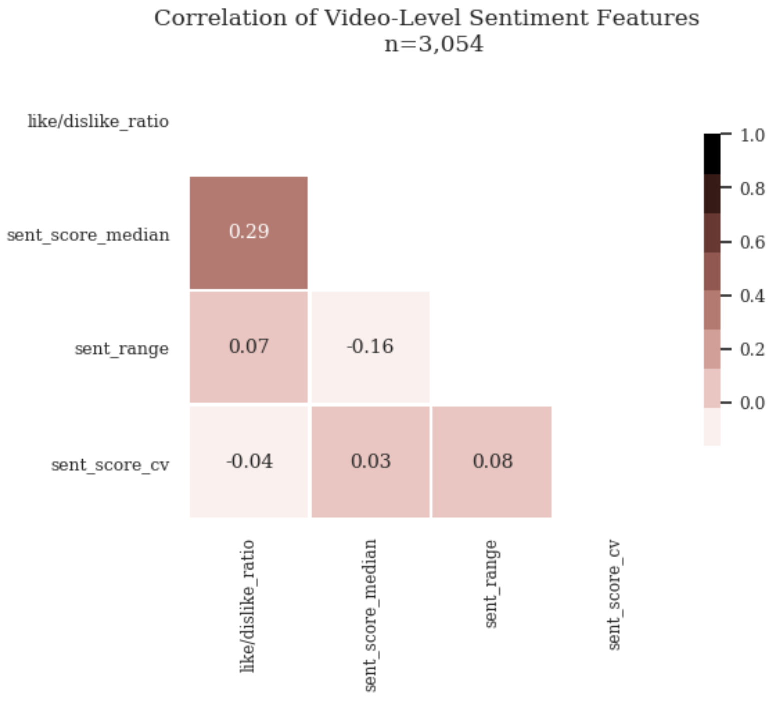


Figure 5. Correlations of video comment sentiment features.

***Question 1 (sentiment)*:**

After outlier removal via Mahalanobis distance, further aggregation and transformation, the video-level derived sentiment features were inspected and modeled to verify their relationships with the like to dislike ratio.

With a Pearson's R of 0.29 shown in in Figure 5 and a p-value that approaches zero when used as a coefficient in an Ordinary Least Squares regression, the null hypothesis can be rejected that the median sentiment of a YouTube video comment section is not related to the like to dislike ratio of the video.

***Question 2 (growth)*:**

All findings are subject to the video sampling methodologies implemented and as such, should be interpreted with the potentiality of bias in the results.

As evidenced in Figure 3, the rate at which viral videos grow tends to decrease very quickly as time progresses. The correlations for the engineered growth features are displayed in Figure 6, where the **viewcount_slope** can be observed to decrease as the video gets older, thus confirming the trend noticed in Figure 3. Moreover, subscriber growth tends to increase almost 1:1 when the channel uploads a new video, implying the importance of frequent uploads for a YouTube channel's continuing success.

Appendix 1 displays correlation coefficients for all videos, not just viral videos. This correlation map includes approximately 1500 videos published in the last 3 years for recency's sake. There are many insights to be extracted from the table in Appendix 1, however, only the most pertinent will be elucidated in pursuit of conciseness. Videos with the category "Music" have the highest positive correlation with **viewcount** meaning that music videos get the most views, which is objectively true. Additionally, "Music" categorized videos have a negative correlation with the **length** variable, implying that music videos are shorter than the rest of the population.

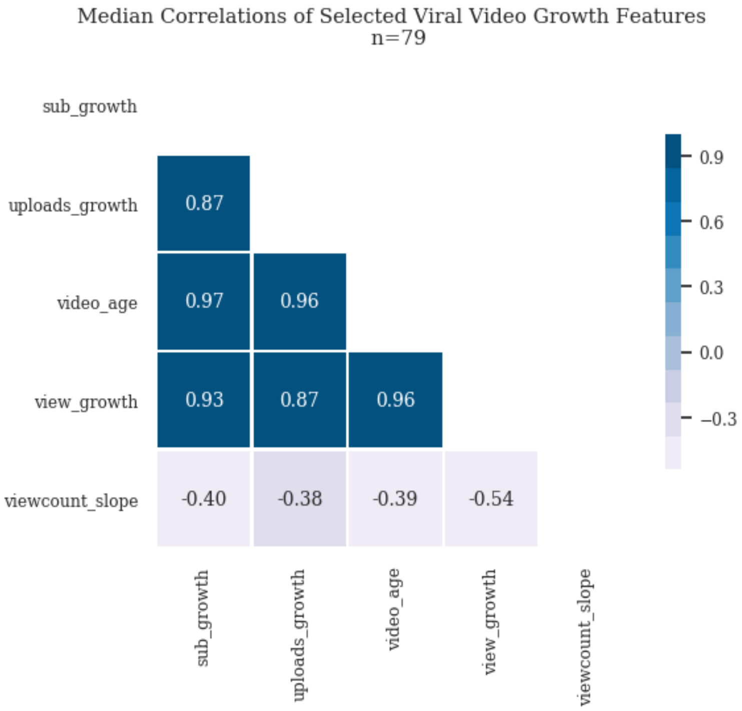


Figure 6. Heatmap of correlations between engineered video growth features.

With respect to the **like/dislike_ratio** feature, it is not surprising that videos with the "News & Politics" category have a negative correlation, as it is very divisive.

More formal, academic videos like those with the categories "Science & Technology", or "News & Politics" have negative correlations with the percentage uppercase in the title, whereas "Gaming" videos have a positive correlation, implying that those videos have more clickbait titles. Interestingly enough, **pct_title_uppercase** is also positively correlated with the **likes/views** ratio, meaning that viewers tend to like a video more when the title is all uppercase. This also may be due to the fact that YouTubers that post all uppercase gaming videos also most likely ask their viewers to like their videos explicitly. This insight is especially relevant to aspiring YouTubers attempting to garner more likes on their videos.

One interesting insight is that channels that post "Music" videos do not have a strong positive correlation with **subscriber_count**. This demonstrates how many people use YouTube to simply listen to music, and don't subscribe to the channels that post the music. Channels that post videos with the categories "Entertainment", "Howto & Style", and "Pets & Animals" tend to have more subscribers; this implies that they post more regularly, which is why a user subscribes (to see more in the future).

Unfortunately, there is no discernable difference in the growth rates when partitioned by video category, save for "Travel & Events", which upon inspection, was anomalous. However, the above insights provide an ample springboard for any budding YouTuber to better understand how videos grow, for how long they grow, the types of videos that receive the most attention, and how to maintain a steadily increasing subscriber base.

References

[1] ‘You know what’s cool? A billion hours’, *Official YouTube Blog*. [Online]. Available: https://youtube.googleblog.com/2017/02/you-know-whats-cool-billion-hours.html. [Accessed: 10-Dec-2019].

[2] J. Hale, ‘More Than 500 Hours Of Content Are Now Being Uploaded To YouTube Every Minute’, *Tubefilter*, 07-May-2019. [Online]. Available: https://www.tubefilter.com/2019/05/07/number-hours-video-uploaded-to-youtube-per-minute/. [Accessed: 12-Dec-2019].

[3] ‘Beautiful Soup Documentation — Beautiful Soup 4.4.0 documentation’. [Online]. Available: https://www.crummy.com/software/BeautifulSoup/bs4/doc/. [Accessed: 10-Dec-2019].

[4] C. Blythe, ‘pseudobrilliant/youtube_analyzer’, 01-Sep-2019. [Online]. Available: https://github.com/pseudobrilliant/youtube_analyzer. [Accessed: 10-Dec-2019].

[5] ‘Pafy Documentation — pafy 0.5.1 documentation’. [Online]. Available: https://pythonhosted.org/pafy/. [Accessed: 10-Dec-2019].

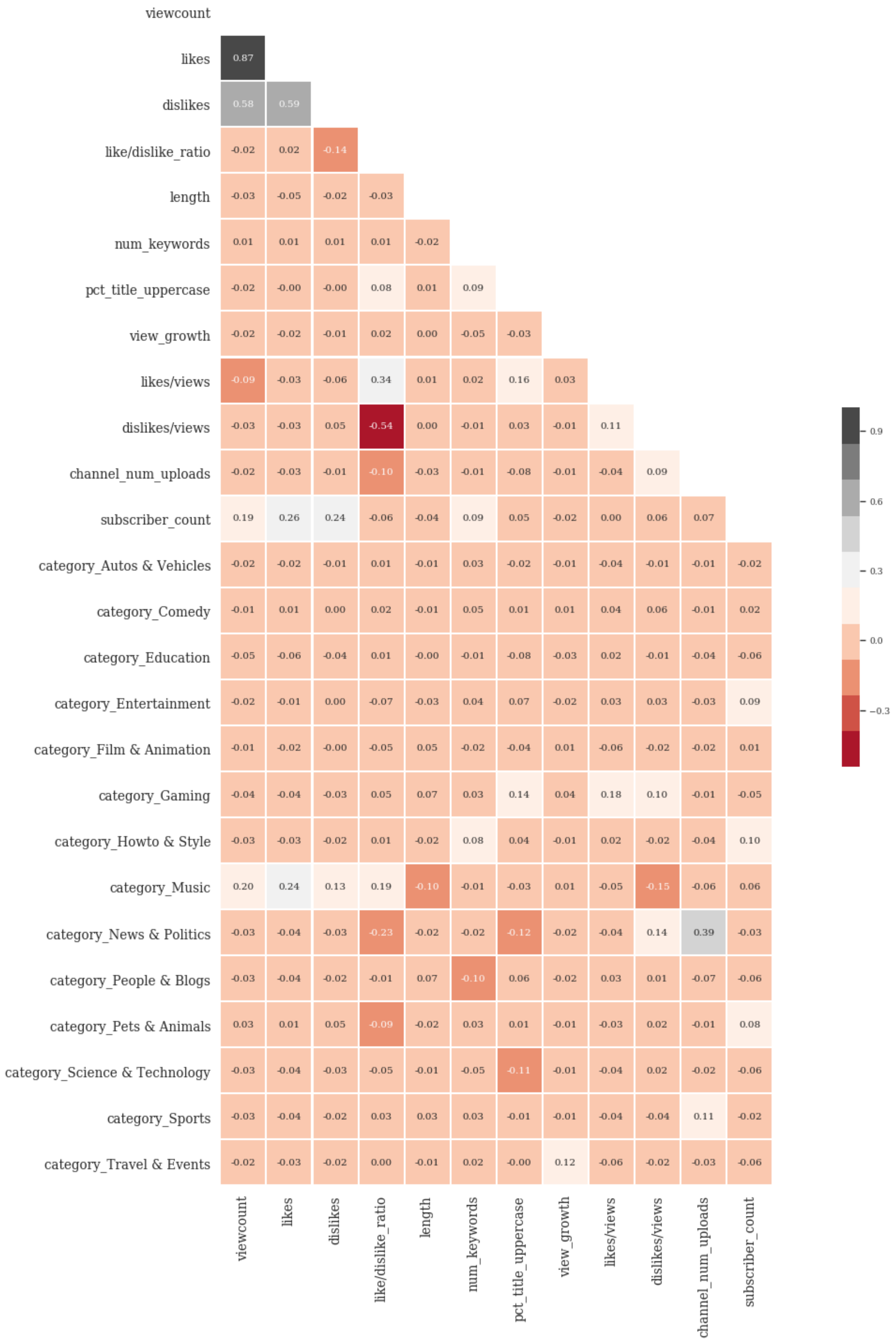
[6] ‘first20hours/google-10000-english’, *GitHub*. [Online]. Available: https://github.com/first20hours/google-10000-english. [Accessed: 10-Dec-2019].

[7] ‘nltk.sentiment package — NLTK 3.4.5 documentation’. [Online]. Available: https://www.nltk.org/api/nltk.sentiment.html. [Accessed: 11-Dec-2019].

[8] ‘numpy.gradient — NumPy v1.17 Manual’. [Online]. Available: https://docs.scipy.org/doc/numpy/reference/generated/numpy.gradient.html. [Accessed: 11-Dec-2019].

Appendix 1.

Correlation Matrix of Selected YouTube Video Characteristics for Videos Published in Last 3 Years
n=1,460



	viewcount	likes	dislikes	like/dislike_ratio	length	num_keywords	pct_title_uppercase	view_growth	likes/views	dislikes/views	channel_num_uploads	subscriber_count
viewcount												
likes	0.87											
dislikes	0.58	0.59										
like/dislike_ratio	-0.02	0.02	-0.14									
length	-0.03	-0.05	-0.02	-0.03								
num_keywords	0.01	0.01	0.01	0.01	-0.02							
pct_title_uppercase	-0.02	-0.00	-0.00	0.08	0.01	0.09						
view_growth	-0.02	-0.02	-0.01	0.02	0.00	-0.05	-0.03					
likes/views	-0.09	-0.03	-0.06	0.34	0.01	0.02	0.16	0.03				
dislikes/views	-0.03	-0.03	0.05	-0.54	0.00	-0.01	0.03	-0.01	0.11			
channel_num_uploads	-0.02	-0.03	-0.01	-0.10	-0.03	-0.01	-0.08	-0.01	-0.04	0.09		
subscriber_count	0.19	0.26	0.24	-0.06	-0.04	0.09	0.05	-0.02	0.00	0.06	0.07	
category_Autos & Vehicles	-0.02	-0.02	-0.01	0.01	-0.01	0.03	-0.02	-0.01	-0.04	-0.01	-0.01	-0.02
category_Comedy	-0.01	0.01	0.00	0.02	-0.01	0.05	0.01	0.01	0.04	0.06	-0.01	0.02
category_Education	-0.05	-0.06	-0.04	0.01	-0.00	-0.01	-0.08	-0.03	0.02	-0.01	-0.04	-0.06
category_Entertainment	-0.02	-0.01	0.00	-0.07	-0.03	0.04	0.07	-0.02	0.03	0.03	-0.03	0.09
category_Film & Animation	-0.01	-0.02	-0.00	-0.05	0.05	-0.02	-0.04	0.01	-0.06	-0.02	-0.02	0.01
category_Gaming	-0.04	-0.04	-0.03	0.05	0.07	0.03	0.14	0.04	0.18	0.10	-0.01	-0.05
category_Howto & Style	-0.03	-0.03	-0.02	0.01	-0.02	0.08	0.04	-0.01	0.02	-0.02	-0.04	0.10
category_Music	0.20	0.24	0.13	0.19	-0.10	-0.01	-0.03	0.01	-0.05	-0.15	-0.06	0.06
category_News & Politics	-0.03	-0.04	-0.03	-0.23	-0.02	-0.02	-0.12	-0.02	-0.04	0.14	0.39	-0.03
category_People & Blogs	-0.03	-0.04	-0.02	-0.01	0.07	-0.10	0.06	-0.02	0.03	0.01	-0.07	-0.06
category_Pets & Animals	0.03	0.01	0.05	-0.09	-0.02	0.03	0.01	-0.01	-0.03	0.02	-0.01	0.08
category_Science & Technology	-0.03	-0.04	-0.03	-0.05	-0.01	-0.05	-0.11	-0.01	-0.04	0.02	-0.02	-0.06
category_Sports	-0.03	-0.04	-0.02	0.03	0.03	0.03	-0.01	-0.01	-0.04	-0.04	0.11	-0.02
category_Travel & Events	-0.02	-0.03	-0.02	0.00	-0.01	0.02	-0.00	0.12	-0.06	-0.02	-0.03	-0.06