

Characteristics of Trending Videos on YouTube

Problem Statement

With a global user base of 2 billion people, YouTube is the second most visited website in the world. It posts an estimated annual revenue of \$16million to \$25billion, a proportion of which is paid to its content creators; that is, users who upload videos to YouTube. Popular videos are given prominence on YouTube's homepage in its "Trending" section, which the company describes as an avenue to "surface videos that a wide range of viewers would find interesting". However, the question remains: which videos are most likely to trend?

To probe YouTube's algorithms, we examine the characteristics of trending videos across several variables, including category, publishing time, views, and likes among other engagement metrics (see Appendix A). Trends are country-specific, and this research concerns itself with trending videos in the United States. We expect such data to be valuable to content creators and advertisers alike.

Primary Research: Sampling Method

We obtained a dataset titled "Trending YouTube Video Statistics" from Kaggle.com. While the list of trending videos on YouTube is updated every 15 minutes, this dataset adopts purposive sampling to compile a list of the top 200 trending videos between 14 November 2017 and 14 June 2018 at 9am daily. To do so in a systematic manner, the creator of this dataset obtained a YouTube Data API and integrated it with a Python script to scrape the relevant data from YouTube.

Methodologies

Data sampling is a tool that guides researchers on "how much data to collect" and "how often it should be collected". To investigate the characteristics of trending videos, the following methodology was employed:

- **Collection period:** Data collection began on November 14th 2017 and ended on June 14th 2018.
- **Sample size:** The scraper collated a daily list over seven months, for a total of 40950 records of trending videos in the United States. The list includes only the top 200 trending videos *daily*, stopping short of an exhaustive list as the latter would be too time consuming and no more accurate nor representative.
- **The frequency of sampling:** The dataset concerns daily trending videos. To identify how the lists changes over time, the Python script scraper was activated to sample trending videos once a day.

Data Validation: Internal and External Validity

While the dedicated YouTube scraper produces a fairly structured dataset, further data-cleaning is necessary to verify the external and internal validity of the data. In this regard, we will first consider the internal validity of the data, with an emphasis on data errors, before investigating its external validity.

Data Errors (Internal Validity)

We identified data errors by systematically checking for "common" data errors:

- **Data Type Check:** "trending_date", which represents the day during which the video was trending, should be converted from "character" to date or timestamp so that we may analyse trending durations.
- **Range check:** Range checks prove especially tricky for this dataset. For example, views range from 549 views to 225million, with a corresponding count of 0 likes and 5million likes respectively. While this casts some doubt on why the first video would be classified as "trending", we can confirm that both are within acceptable range as there is no hard limit on YouTube views: its most-viewed video has 6.6billion views as of January 2020. In what appears to be a YouTube glitch, however, a particular video has more "likes", "dislikes" and comments than views impossible by definition. Hence, this entry was removed.
- **Consistency check:** All videos were published before they were trending. However, there are 6351 unique video ids in contrast to 6455 unique video titles, which suggests the presence of identical videos with different titles. Hence, our analysis will deal with unique video_ids rather than titles.
- **Cross validation:** Ideally, cross-validation should have been performed on a daily basis as the data was being collected. For example, results from the Scraper could be randomly sampled and contrasted against videos on YouTube's Trending page. At this point, however, we have no sources for cross-validation as the list is updated multiple times daily, and views on videos are cumulative with no historical records.
- **Outlier detection:** While variables like "views" and "likes" present large variation and range, we advise against removing "outliers" as doing so would present a distorted picture of which videos make the cut for YouTube's "trending" criteria. Instead, we will explore the entire range (min and max views) to understand the characteristics of trending videos.

58 **Confounding Factors (Internal Validity):** The goal of this research is to understand the characteristics of trending
 59 videos so as to extract recommendations for content publishers (“YouTubers”). However, while examining the
 60 characteristics of trending videos may serve as a proxy measure, we cannot make any claims of causal
 61 relationships, nor can we study all variables exhaustively and determine their respective weights.
 62

63 **External Validity:** This research investigates trending videos in the United States at 9am daily. Hence, while
 64 providing some descriptive insight into the characteristics of trending videos, we are unable to generalize these
 65 results to other times of the day, nor to other countries.
 66

67 **Analysis Approach and Key Assumptions**

68 **Trending Criteria:**

- 69 • **Fair Mix of Variables:** According to YouTube Search Engine metrics published by YouTube Creators,
 70 trending aims to combine “popularity” with “novelty”; and considers engagement metrics such as likes
 71 and comments in addition to views. Hence, the video with the highest number of views may not
 72 necessarily trend at first place; and videos with less than 1000 views may still trend. YouTube does not
 73 disclose the weight of each variable in its algorithm, only clarifying that “trending” is not a space for paid
 74 ads. Hence, this research assumes that YouTube’s algorithms are *fair* in measuring the trending eligibility
 75 of videos. If there is any built-in bias (e.g. sponsored content), all analysis will be equally biased.
 76
- 77 • **Baseline criteria:** In computing the average, percentile ranks, and range for “likes” and “views”, we hope
 78 to (1) understand the most popular categories on YouTube and (2) distil the baseline criteria for trending
 79 videos, or the number of views required to trend. For example, trending videos within Entertainment
 80 have a minimum of just 798 views, giving content creators an idea of the bare minimum criteria to trend.
 81
- 82 • **Application of Criteria at Every Refresh:** YouTube’s trending lists are refreshed every 15 minutes. We
 83 assume that each refresh is independent of the list that precedes it: that is, videos that are already
 84 trending do not have an advantage over videos that are not trending, as both are again subject to the
 85 same evaluation criteria, every 15 minutes, to determine its inclusion on the trending list. Hence, even if
 86 the video_id is repeated, this research treats them as independent entities.

87 **Time:** As the scraper was only activated once daily at 9am, this research assumes that no significant differences
 88 exist in Trending videos across different times of the day. If, however, YouTube alters its algorithm across the day
 89 for different audience compositions—for example, if the 9am trending lists are catered for adults-- then this
 90 research cannot be said to be generalizable of which videos are most likely to trend.
 91

92 **Table 1: Identifying the Top Trending Categories on YouTube**

Row Labels	Count of video_id	% of video ID	Average Views	Max Views	Min. Views
Entertainment	9964	24.33%	2,067,883	149,376,127	798
Music	6472	15.81%	6,201,003	225,211,923	1,591
How to and Style	4146	10.12%	983,730	54,155,921	1,107
Comedy	3457	8.44%	1,480,308	29,178,096	1,807
People and Blogs	3210	7.84%	1,531,835	56,111,957	884
News and Politics	2487	6.07%	592,588	10,277,358	549
Science and Technology	2401	5.86%	1,452,627	42,799,458	983
Film and Animation	2345	5.73%	3,106,250	54,863,912	943
Sports	2174	5.31%	2,025,969	29,090,799	658
Education	1656	4.04%	712,941	7,349,435	773
Pets and Animals	920	2.25%	831,143	6,187,457	3,393
Gaming	817	2.00%	2,620,831	16,935,442	1,237
Travel and Events	402	0.98%	854,620	23,932,421	789
Autos and Vehicles	384	0.94%	1,355,965	25,244,097	2,860
Non-profits and Activism	57	0.14%	2,963,884	24,286,474	1,456
Shows	57	0.14%	903,527	1,445,949	36,609
Overall total	40949	100.00%	2,067,883	149,376,127	798

93 Of 43 possible categories, the top five categories are “Entertainment”, “Music”, “How to and Style”, “Comedy”,
 94 “People and Blogs”, making up 66.54% of all trending videos. In fact, only 15 categories appeared within the
 95 trending list, with zero representation from categories like “foreign”, “horror”, or “documentary”. While those are
 96 not indie categories, it suggests that some categories are significantly more likely to trend than others, and
 97 YouTube does not make provisions for equal representation across categories.

98
 99 While trending videos have 2.4million views on average, this figure is distorted by viral videos with view counts as
 100 high as 225million. Interestingly, the minimum values for views suggest that videos with less than 1000 views are
 101 eligible to trend too, which confirms that additional variables (e.g. channel, likes, and referral source) are at play
 102 in determining the list of trending videos.

104 **Table 2: Publishing Day and Time**

	12AM - 4AM	4AM - 8AM	8AM - 12PM	12PM - 4PM	4PM - 8PM	8PM - 12AM	Grand Total
Monday	2.54%	1.06%	1.27%	4.63%	4.00%	1.58%	15.08%
Tuesday	2.07%	0.87%	1.57%	5.27%	4.42%	2.38%	16.57%
Wednesday	2.44%	1.11%	1.31%	4.08%	4.51%	3.06%	16.51%
Thursday	2.23%	1.39%	1.59%	5.95%	3.79%	2.02%	16.97%
Friday	2.70%	2.50%	1.82%	4.14%	3.72%	2.21%	17.10%
Saturday	1.38%	0.41%	0.44%	2.55%	2.90%	1.09%	8.77%
Sunday	1.04%	0.65%	0.67%	2.84%	2.54%	1.24%	8.98%
Grand Total	14.40%	8.00%	8.68%	29.46%	25.90%	13.57%	100.00%

105
 106 Publishing day and time appears to be a significant variable in determining trending likelihood. In particular,
 107 videos published between 12pm to 4pm on Thursdays are most likely to trend: with 5.95% representation of
 108 the overall total, its trending probability is 14 times more than videos published on Saturdays, between 4am
 109 to 8am (for which the probability is lowest). In fact, videos published on weekends are at least *half* as likely
 110 to trend. Hence, content creators who wish for exposure on YouTube’s homepage should aim to publish
 111 their videos on weekdays—especially Fridays or Thursdays—between 12pm to 4pm or 4pm to 8pm.

112
 113 **Table 3: Engagement Metrics on Trending Videos**

114 As YouTube clarifies, its trending criteria considers more than just views to include audience engagement
 115 metrics such as likes, dislikes, and comments as well. Hence, we had expected to find that videos within the
 116 top five trending categories would present:

- 117 - **Higher likes/dislikes ratio**, indicating that likes significantly outweigh dislikes.
- 118 - **Higher likes/views ratio** and **higher comments/views ratio**, indicating that these videos sufficiently
 119 motivate audiences to engage with a “like” or a comment.

Category by frequency	Engagements on Average				Engagements Ratio		
	Likes	Dislikes	Comment	Views	Likes / Dislikes	Likes / Views	Comment / Views
Top 5							
Entertainment (f = 24.3%)	53,243 (92.5%)	4,314 (7.5%)	7,383	2,067,883	12.3	2.57%	0.36%
Music (f = 15.8%)	218,918 (96.51%)	7,908 (3.49%)	19,360	6,201,003	27.7	3.53%	0.31%
How-to and Style (f = 10.12%)	39,286 (96.75%)	1,320 (3.25%)	5,584	983,730	29.8	3.99%	0.57%
Comedy (f = 8.44%)	62,582 (96.77%)	2,092 (3.23%)	6,522	1,480,308	29.9	4.23%	0.44%
People and Bogs (f = 7.84%)	58,136 (94.82%)	3,174 (5.18%)	7,719	1,531,835	18.3	3.80%	0.50%
Total (Top 5 Categories)	432,166 (95.83%)	18,808 (4.17%)	46,567	12,264,760	23.0	3.52%	0.38%

Bottom 5							
Non-profits & Activism (f = 0.14%)	259,924 (81.74%)	58,077 (18.26%)	84,365	2,963,884	4.5	8.77%	2.85%
Shows (f = 0.14%)	18,994 (97.79%)	430 (2.21%)	1,669	903,527	44.2	2.10%	0.18%
Auto & Vehicles (f = 0.94%)	11,056 (94.58%)	633 (5.42%)	2,043	1,355,965	17.5	0.82%	0.15%
Travel & Events (f = 0.98%)	12,030 (93.42%)	847 (5.42%)	2,267	854,620	14.2	1.41%	0.27%
Gaming (f = 2.0%)	84,502 (88.26%)	11,242 (6.58%)	18,042	2,620,831	7.5	3.22%	0.69%
Total (Bottom 5 Categories)	386,506 (84.44%)	71,228 (15.56%)	108,386	8,698,827	5.4	4.44%	1.25%

121

122

123

124

125

126

127

128

129

130

131

132

133

134

135

136

137

138

139

We find that trending videos within the top five categories are indeed more likely to present a higher likes/dislikes ratio (23.0), which is more than four times higher than that for the bottom five categories (5.4). This confirms our hypothesis that videos with more likes—or less dislikes—are more likely to trend.

However, the bottom five videos present higher likes/views ratio and higher comments/views ratio, which challenges our hypotheses that videos with greater engagement on these specific metrics are more likely to trend. Upon closer analysis, it appears that these ratios are balanced against the likes/dislikes ratio: for example, while “Non-profits and Activism” presents the highest likes/views ratio and comment/views ratio, it is also least likely to trend—possibly because videos tend to be controversial, presenting the lowest likes/dislikes ratio of 5.4.

We also find mixed results within individual categories, reflecting the sophistication of YouTube’s algorithms. For example, “Entertainment” presents the second smallest likes/dislikes ratio of 12.3, reflecting mixed opinions and a high propensity for audiences to dislike the video; whilst “shows” have the highest likes/dislikes ratio of 44.2. However, “Entertainment” videos have the highest representation (24.3%) in Trends, while “Shows” comprise just 0.14% of trending videos. We may thus conclude that videos with more likes are more likely to trend, but it is not considered in isolation and there are other variables with greater weights—including view count, video category, and publishing time—that affect trending probability.

140
141

Appendix A: Variables in the Dataset

Variables	Data Type	Remarks
video_id	character	Each video is given a unique video ID. If a video trends for multiple days, the video_id will appear multiple times.
trending_date	numeric	Data-cleaning: to convert to date. Indicates that the video was trending at the point of data collection.
title	character	Title of YouTube Video
channel_title	character	Username of Content Uploader
category_id	character	44 categories possible (e.g. entertainment, news and politics)
publish_time	date	Time at which the video was published
tags	character	Users add tags to video to facilitate YouTube searches
views	numeric	View count (cumulative) at the point of data collection
likes	numeric	Likes (cumulative) at the point of data collection
dislikes	numeric	Dislikes (cumulative) at the point of data collection
comment_count	numeric	No. of comments (cumulative) at the point of data collection
thumbnail_link	character (link)	Link to a picture of a video thumbnail
comments_disabled	binary	True/False: If comments were disabled
ratings_disabled	binary	True/False: If ratings were disabled
video_error_or_removed	binary	True/False: If a video was deleted
description	character	User-generated video descriptions

142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158

References

- Google. (n.d.). Trending on YouTube - YouTube Help. Retrieved January 29, 2020, from <https://support.google.com/youtube/answer/7239739?hl=en>
- Kaggle. (2017, November 14). Trending YouTube Video Statistics. Retrieved January 27, 2020, from <https://www.kaggle.com/datasnaek/youtube-new>
- GitHub. (2018, November 1). Trending-YouTube-Scraper. Retrieved January 29, 2020, from <https://github.com/mitchelljy/Trending-YouTube-Scraper/blob/master/scraper.py>
- Kaggle. (2018, February 23). Older Data. Retrieved January 29, 2020, from <https://www.kaggle.com/datasnaek/youtube-new/discussion/50399>
- YouTube. (2017, September 1). How YouTube's Trending Tab Works. Retrieved January 27, 2020, from <https://www.youtube.com/watch?v=GZmGmkOJ9ME>