Viewed More

Measuring the effect of YouTube video title wording on view count

Erik Kessler and Greg Szumel
Williams College
CSCI 375
Spring 2017

May 17, 2017

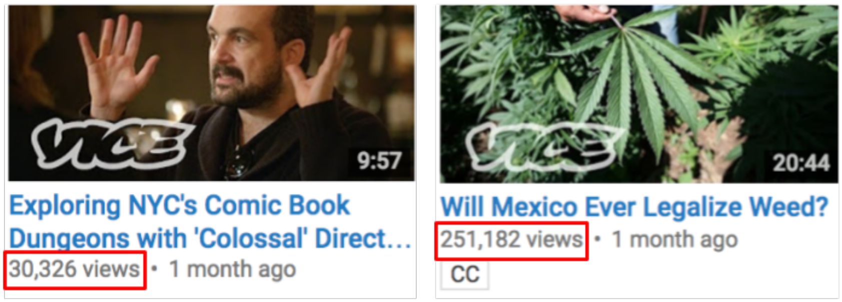
Abstract

We created a Naive Bayes classifier designed to predict whether a YouTube video is likely to receive more or less views than average for the uploader based on the title. Such a system would allow YouTube content creators to increase their view counts and would allow YouTube to increase traffic and user engagement. We created a dataset of 25k video titles from the top 1600 subscribed YouTubers. Our dataset is labeled based on the number of standard deviations away from the uploader's mean view count per video the video is. We created a variety of features including ***. We were able to achieve ***, and our classifier works best when ***.

1 Introduction

Every minute, hundreds of hours of video are uploaded to YouTube. How can YouTube content creators or uploaders stand out from the crowd and draw an audience to their videos? One tool they have at their disposal is the video title. Video creators try to make their videos sound enticing by putting a flashy and interesting title to spark interest and curiosity in users with the hope the user will choose to watch their video. View counts are not only a measure of popularity but are also the main currency of YouTube as more views correspond to more opportunities for advertisements. Therefore, video creators would be interested in improving their titles to earn more money and draw more subscribers to their channel. YouTube itself is interested in keeping users on their site by making sure users continue to see interesting titles and continue to click through to those videos.

There would be value for both YouTube and video uploaders alike if when a creator uploaded a video, the site could indicate the quality of the title and even suggest a better title that would draw more views. The goal of this project was to develop a system that could predict how a video would do based on different features of the title. Such a system could also be used to filter interesting titles towards the top of search results or towards the top of a suggested videos page. Furthermore, this could be extended to detect video titles that go to far and cross into the category of clickbait which YouTube might want to filter out to avoid disappointing and annoying users.



Exploring NYC's Comic Book Dungeons with 'Colossal' Direct...
30,326 views • 1 month ago

Will Mexico Ever Legalize Weed?
251,182 views • 1 month ago

Figure 1: Two videos from the same uploader with drastically different view counts.

2 Related Work

Our interest in this project was initially sparked by Chenhao Tan, Lillian Lee, and Bo Pang's project *Retweeted More* [3]. In *Retweeted More*, they developed an algorithm for predicting which of two tweets on the same topic would receive more retweets. They found that wording does matter. There are features that improve the probability that a tweet will be retweeted. This finding suggests that we will find features in YouTube title wordings that influence the number of views. We also predict that the features that improve YouTube titles will differ from those that improve tweets which is why we were interested in expanding on that project to explore YouTube. Furthermore, for this existing work, creating a dataset of content and author controlled tweets while also controlling for other confounding effects such as when the tweets were sent was a major component of the project and as they creatively used tweets-pairs from the same account linking to the same content but with different wordings as natural experiments. We would need to find or create a similar dataset for YouTube videos.

Existing work on determining which factors impact the popularity of YouTube videos have focused on factors such as social networks and view count histories [1]. It is difficult for a YouTuber to easily change the structure of their social network or their video's viewing patterns, so we wanted to focus on something that the uploader could easily adjust that would help them achieve more video views. Since we knew from Tan that wording had an impact on retweets and Borghol did not look at title wording, we realized we could add to existing research by studying the effect of title wording on view counts. Borghol also offered us a promising dataset of cloned videos which were sets of different re-uploads of the same video by different users. Unfortunately, many of the videos were non-English and those that were in English were movie trailers so they did not capture the types of videos on YouTube we wanted to explore.

Another piece of relevant work we looked at was Himabindu Lakkaraju's work on examining the effect of titles on the success of Reddit image posts [2]. Their finding that the title wording influences the success of a post further confirmed that wording matters, but research is missing on title wording on YouTube and what features make a good video title. Additionally, this work provides some starting points as the authors identify some linguistic features of titles that impact popularity.

Overall, out survey of existing work shows that wording matters, having a dataset that controls for possible confounding variables is important, and there is a current lack of literature on what features of YouTube video titles predict more popularity and views. We look to build off this existing work with our own study of wording effects on YouTube titles.

3 Formulation

We formulate our problem as a classification problem. We are looking to classify a string of words representing a title as either a "good" or "bad" title where good means the title will likely get more than average views and bad means the title will likely get fewer views than average.

To do the classification, we will construct a feature vector for the title using a variety of features that we think would be useful for predicting how a title would perform. We can then use a set of labeled feature vectors to train a classifier and use that classifier to predict the quality of unseen titles.

4 Architecture Overview

Provide an overview of your system, presenting and justifying important design choices such as algorithms, features, etc.

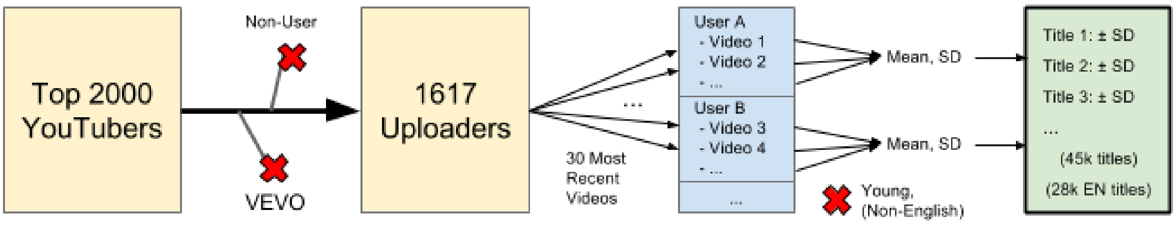


Figure 2: Diagram showing the dataset creation process resulting in 45k total titles and 28k English titles.

Universal			
Use	*SD Thresh.*	*GOOD*	*Total*
Test	0.00	4741	13625
Train	0.00	11148	31789
Train	0.25	8533	24812
Train	0.50	6652	15725

English-Only			
Use	*SD Thresh.*	*GOOD*	*Total*
Test	0.00	2869	8400
Train	0.00	6631	19598
Train	0.25	5037	15182
Train	0.50	3899	9360

Table 1: Table showing statistics on the dataset.

5 Experiments & Analysis

Describe the dataset and the experiment setup. Compare the performance of your system to that of a related work or other versions of your system (both if possible!). As part of the analysis, examine the specific data points that are correctly classified by one but not others, instead of simply focusing on the final performance scores.

6 Limitations & Future Work

Summarize the project (major findings, weaknesses, etc.) and describe an extension of the project that youd like to try if you had more time.

Future Work: Use descriptions, clickbait detection, larger dataset for even better CNN performance, different communities expect different things, title generation

7 Conclusion

References

[1] BORGHOL, Y., ARDON, S., CARLSSON, N., EAGER, D., AND MAHANTI, A. The untold story of the clones: Content-agnostic factors that impact youtube video popularity. In *Proceedings of KDD* (2012).

[2] LAKKARAJU, H., MCAULEY, J., AND LESKOVEC, J. Whats in a name? understanding the interplay between titles, content, and communities in social media. In *Proceedings of ICWSM* (2013).

[3] TAN, C., LEE, L., AND PANG, B. The effect of wording on message propagation: Topic- and author-controlled natural experiments on twitter. In *Proceedings of ACL* (2014).

A Roles

Describe the roles of group members.