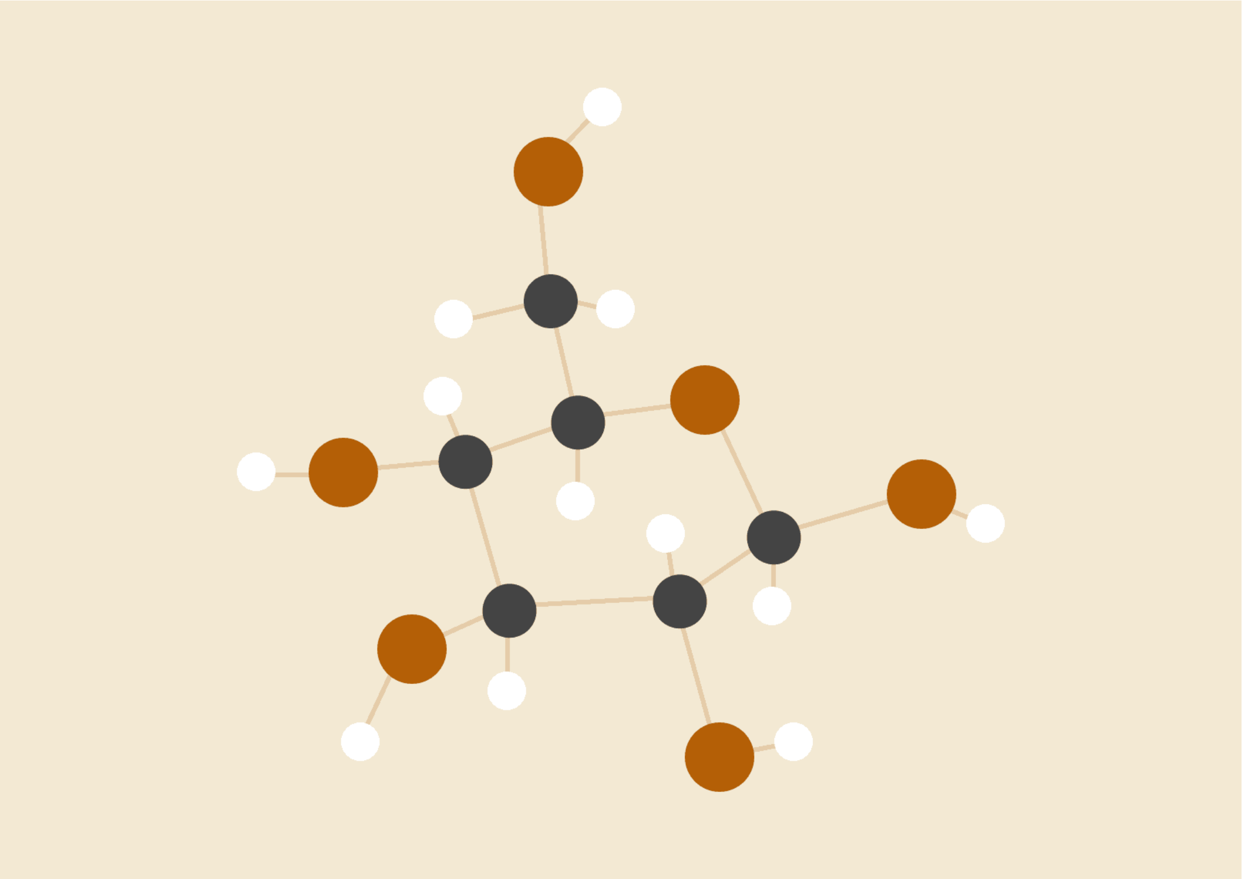
Online Engagement on Ted Talk Videos

Using Machine Learning Techniques to Study Online Content Engagement



Nikhil Nandigam and Jesse Velasquez

Fall 2021

IMT 574, UW Information School

INTRODUCTION

In the course of this report, we will use machine learning techniques to understand how online content creators can boost audience engagement. In general, video engagement is considered “the primary factor that determines how much information viewers will retain from the videos they watch'' with higher engagement being desirable (Iserovitch, 2021). In the age of social media marketing and influencers, content creators seek to grow engagement with their audience to foster a lasting connection with their target demographic and demonstrate their ability to drive market interest toward the brands or products sponsoring their content. Engagement comes in various forms depending on the capabilities and goals of each online platform. In our analysis, we will primarily focus on viewers’ first level comments and number of views as an indicator of engagement.

This study uses the TED Main dataset containing metadata about 2,550 talks uploaded to the official TED.com website until September 21, 2017. The data includes features answering questions such as who delivered the talk, what topics the talk covered, where the talk was delivered, when the speaker delivered the talk as well as when the video was uploaded, and how online viewers viewed, commented, and rated the talk’s video. Meanwhile, the TED Transcripts dataset contains the official English-language transcript of 2,467 talks alongside their respective URL. This data has been scraped from the official TED.com website and is available under the Creative Commons License. Both datasets were joined by the URL variable and cleaned by removing rows containing missing transcript values.

TED has stated that their “agenda is to make great ideas accessible and spark conversation” (TED, n.d.). To maximize this goal, TED Talks must be highly engaging – in our study, we examine how this can be done through the following questions:

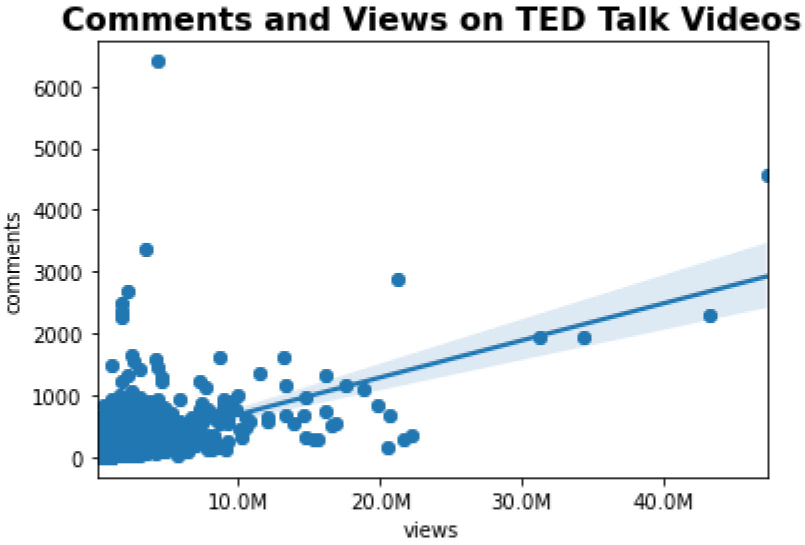
- Do the number of views correlate to engagement?
- How can we use video attributes to predict engagement?
- What attributes result in high audience engagement?

DO THE NUMBER OF VIEWS CORRELATE TO ENGAGEMENT?

As a starting point for our research into boosting online engagement, we test the logical assumption that the more views a piece of content receives, the more engagement it will create among viewers. Furthermore, we want to learn how much of the variance in engagement by way of number of comments can be explained by predictors such as views -- but also, the duration of the video, multilingual availability, and the number of speakers the piece of content features.

To do this, we run a linear regression model. With the number of views as our sole predictor, our model returns an R-squared value of 0.461. With the additional three predictors (duration, multilingual availability, number of speakers), the R-squared jumps to 0.535. Therefore, it seems the number of views accounts for around 46% of the variance we see in the engagement.

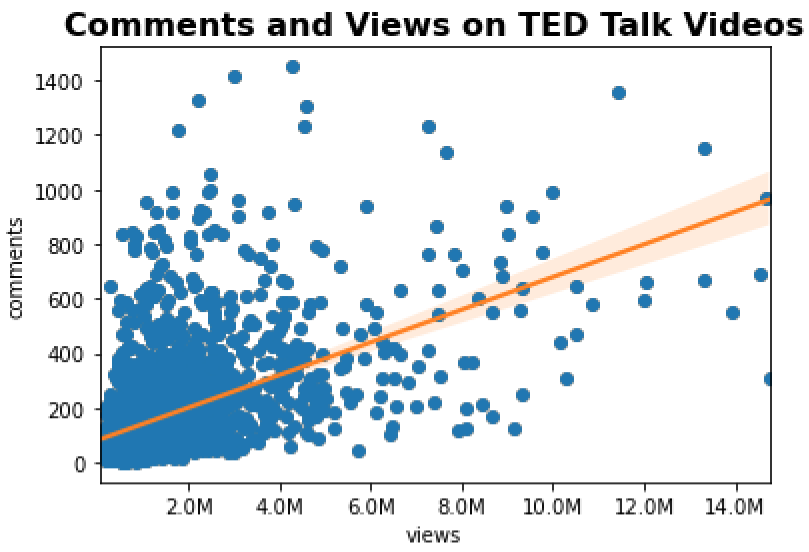
The plot below depicts each talk's views on the x-axis with the number of comments on the y-axis, along with the linear regression line. The positive coefficient indicates a positive relationship between the predictor and label variables.



Linear equation: comments = 0.00007688(views)
No interpretable constant in this context
R-squared = 0.461

This plot shows many outliers, both in terms of the number of views as well as the number of comments received. By filtering some extreme outliers, we can learn more about the majority of the TED talk videos concentrated on the bottom-left corner of the plot.

Filtering extreme outliers by examining those videos with 15 million or fewer views and 1,500 or fewer comments results in losing 30 of 2,550 observations. Plotting the remaining 2,520 points and running a linear regression model increases our R-squared value to 0.566 with views as our sole predictor, and 0.65 with all four predictors (views, duration, multilingual availability, number of speakers).



Linear equation: comments = 0.0000854(views)
No interpretable constant for this context
R-squared = 0.566

Although we can continue to increase the model's accuracy by further cutting down on outliers, this represents a design choice and would require further domain knowledge to best configure bounds for outliers. Filtering outliers comes at a cost as we do not want to build a model that performs well on training data, but falls victim to overfitting when confronting new data.

HOW CAN WE USE VIDEO ATTRIBUTES TO PREDICT ENGAGEMENT?

Although the linear regression model partially explains the variance in the correlation between views and comments, we can conceptualize audience engagement with more complexity as a ratio of comments to views.

Engagement = Comments / Views

Furthermore, we use this definition of engagement to create five classes. We can call these classes Very Low Engagement, Low Engagement, Medium Engagement, High Engagement, Very High Engagement -- each represents a particular range of engagement scores, thereby allowing us to transform the continuous engagement value into a discrete one that we can use for classification. We experimented with various numbers of classes and found a steep dropoff in model accuracy beyond five, so using five classes is how we proceeded.

With these five levels of engagement scores serving as our label, we can then deploy a Naïve Bayes technique to test how accurately some key string variables perform in predicting how much engagement a particular video might receive. Naïve Bayes is deliberately chosen because of its versatility and speed in handling large volumes of text data and its proven ability to perform well under multiclass prediction. Considering the large volume generated from our tokenized strings of the transcript variable alone, these model traits are ideal for our application. Other methods such as random forest and KNN were considered but found to be error-ridden and immensely cost prohibitive in comparison. Of the various Naïve Bayes classifiers available, we used the multinomial approach because it is most suitable for classifying discrete features, such as the word counts from our tokenized strings, and this was found to be more accurate than the Gaussian approach.

The model reports the following average accuracy scores across ten runs, each using a random 80/20 train-test split in the data. A larger number of models was initially considered, but after discovering that nearly all model accuracies were found to have a standard deviation of 0 or nearly 0, they were reduced to mitigate costs without sacrificing much accuracy.

PREDICTOR	AVERAGE ACCURACY	PROCESSING TIME (seconds)
Tags	0.873	0.435
Title	0.914	0.371
Description	0.911	1.110
Transcript	0.915	28.063

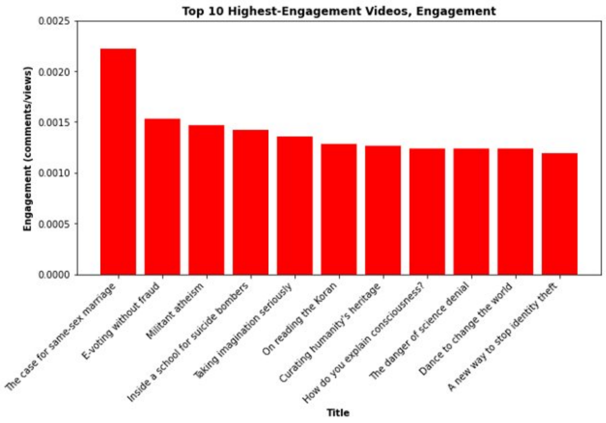
The results of the Naïve Bayes model demonstrates that although the entire English-language transcript serves as the best predictor for the range of engagement a video is likely to receive, it may not be the most practical of the predictors. It would be more computationally efficient to use the video title given the relatively similar accuracy score, but much faster processing time.

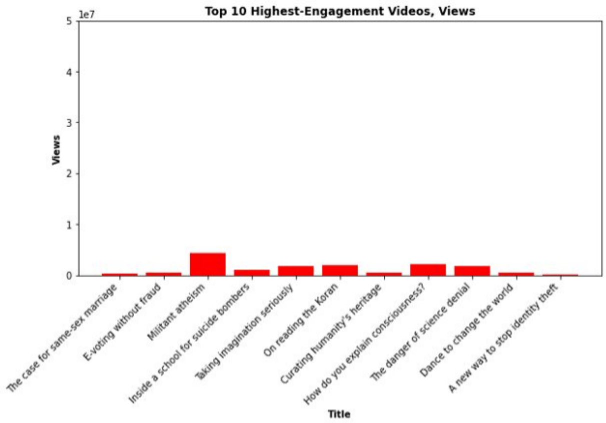
Given the practicality of using the video titles to predict audience engagement, we will proceed to more closely examine the titles of videos with the highest engagement scores.

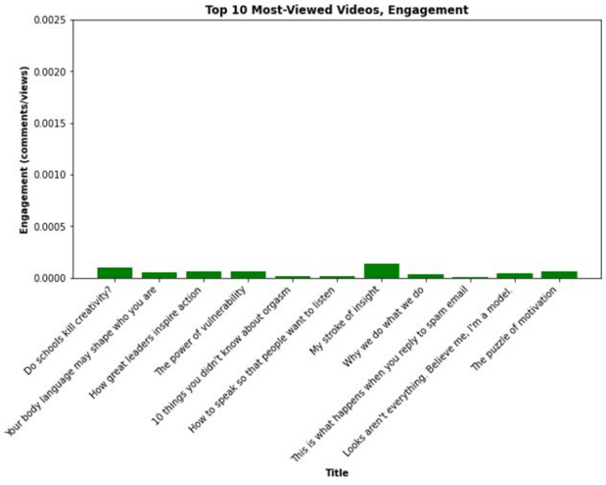
WHAT ATTRIBUTES RESULT IN HIGH AUDIENCE ENGAGEMENT?

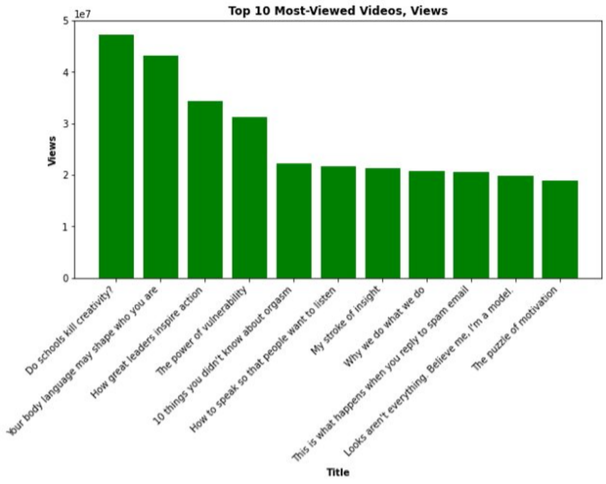
When we consider videos with the highest engagement and those with the highest views, we notice a pattern. Almost all high-engagement videos, including “The case for same-sex marriage,” “Militant atheism,” and “The danger of science denial,” have titles pertaining to topics of controversy, contention, or otherwise inducing of strong emotion.

Meanwhile, keywords in the titles of videos with the most views, including “How great leaders inspire action,” “The power of vulnerability,” or “My stroke of insight,” are associated with positivity. The most attractive TED Talks are the ones where people experience the most catharsis or empowerment. Interestingly, there does not seem to be any overlap between the two extremes. The highest-engagement videos are certainly not the most viewed. Likewise, the most-viewed videos do not have a high level of engagement, as demonstrated below.









Based on this visualization, high engagement alone is not a good thing without qualifiers. This parallels what we already understand to be true of social media. Videos leveraging a strong emotional response, usually with a polarizing topic, tend to drive the most interaction with their audience. TED can use this evidence to boost engagement per their mission statement, but it's a dangerous tactic because it can enable harmful rhetoric to gain more traction, and they must therefore strike a balance between the empowerment of their popular videos and the emotion of the engaging videos.

What an impactful video requires is some sort of maximum *positive engagement*, which could be defined as a combination of high engagement and positive ratings. With each video, one would have to extract the top ratings (values such as "inspiring" or "funny") and flag each word with a designation of positive or negative, then perhaps generate a positivity score to combine with engagement. Several attempts were made to achieve this, but it was concluded that the complexity of the language processing and analysis required was both very subjective and beyond the scope of our project, so we did not pursue it further. However, it should be considered a natural next step.

RECOMMENDATIONS TO CONTENT CREATORS

Views are important to generate engagement, but they do not represent the entire picture. While content creators may logically look to grow their views in order to drive engagement, our linear model shows that although views can explain a chunk of the variance in engagement as defined by the number of comments on a video, it may not be the best determinant for creators to target. Rather, content creators must take a broader approach. In addition to factoring in the length and multilingual reach of their content, they must consider the topics they discuss, the keywords they use, and the timeliness of these conversations.

Instead of fussing over viewer or subscriber metrics, consider other subject-matter optimization techniques. While creators often chase view counts, our models suggest that the targeted optimization of the subject-matter of video content may prove a more fruitful and efficient path to boosting audience engagement. Content creators may consult with tools such as search engine optimization or demographic data analytics to determine which keywords best engage the audience they seek to attract.

Engage emotions to drive engagement. Once conducting optimization techniques to attract viewers to their content, creators must ensure that the content itself sparks an emotional connection to as many viewers as possible. Maximizing the level of this impact is what separates the high-engagement videos from the most-viewed, but low-engagement content. Creators may steer this emotional connection by assessing the timeliness and ongoing debate around the topics they address.

ETHICAL ISSUES AND SOCIAL IMPACT

The results presented using these machine learning techniques lend themselves to a conversation on the externalities of purposefully amplifying the dissemination and engagement of online content.

As identified in our data, many of the TED talks which received the highest levels of engagement were those which discussed particularly contentious issues such as religion or other divisive social issues. While TED curates its speakers before providing them with a platform, most other social media platforms do not. When platforms and content creators have a financial incentive to drive views or engagement, it becomes easy to sacrifice vetted, authoritative content for an algorithm prioritizing content with incendiary keywords or otherwise search engine optimized features. This process can quickly become further muddled with corporate and political interests as online platforms use paid advertising to support their hosting platforms.

Therefore, we hope this analysis can be a part of the ongoing discussion over the role video hosting platforms play in our society. Do these platforms represent a private corporate entity or do they serve as public utilities requiring government oversight and regulation? Furthermore, what role does the IT community play to ensure society is literate in the impact of these technologies on which we have become so dependent, especially as we observe their ability to influence our democracies and pandemic response worldwide?

REFERENCES

Iserovitch, T. (January 8, 2021). What is Video Engagement and how to Measure it. *Cincopa*. https://www.cincopa.com/blog/what-is-video-engagement-and-how-to-measure-it/

TED. (n.d.). *Our organization*. https://www.ted.com/about/our-organization