Hector Nevarez
Ling 571

Understanding the Language Used in YouTube Science & Technology Videos

I. Introduction

This report explores the different aspects of the language in science & technology YouTube video titles. The objective of this analysis is to better understand the relationship between the number of views on a video and its direct correlation with the language used in the title. Through this exploration, I will provide insight as to what type of language yields the most viewers. This analysis could be used by current content creators on the platform looking to grow their channel or aspiring content creators, like myself, who are looking to begin and develop a YouTube channel.

II. Collecting Data

YouTube makes the data collection process difficult and time consuming. My goal with the data collection process was to obtain a representative sample of YouTube video titles so I could create a corpus of video titles. However, YouTube is a large platform having over 500 hours of videos uploaded per minute [1]. I designed a couple of parameters for the type of data I would scrape from YouTube. First, the videos I collected would have to be from the science & technology genre. Every YouTube genre has their own type of language and because of the limited resources I had available, focusing on one community of videos would allow for better insight. I also made sure all the video titles were in English, narrowing my data to English science & technology YouTube videos. Lastly, I made sure to not include viral or trending videos. The discrepancy between the views of a viral video compared to an average video were so large that my data would have become skewed. Ultimately, I wanted to collect a representative sample and any outliers would affect my small dataset.

To collect the data, I first needed a list of YouTubers that made science & technology videos. I used the YouTube Channels Kaggle Dataset [2] to gather a collection of science & technology YouTubers. By accessing the data frame and filtering the category name to only science & technology, I was able to gather a list of 3,483 YouTube channels. I broke the data collection down by first populating a csv file with different videos from each channel and then collecting metadata from each video.

To collect the different videos from each YouTuber, I used the Python package selenium to scrape the channels video feed [3]. The Python script would scroll down the channels feed, loading more videos by the user. After automatically scrolling, it would grab the links to all the videos that were loaded in and save them to a file. This process would repeat for every channel in the data frame. Time was the biggest challenge with this process as it took around 8 hours to collect every channels video information.

After collecting a list of YouTube videos from different creators, a second Python script would load every video extracting the video’s title as well as other metadata such as number of views, likes, dislikes, and the description. However, to collect data on the tens of thousands of video links would have taken over 25 hours. For this report I only collected data on 9,000 different videos.

III. Cleaning the Data

As mentioned in the previous section, I had set parameters as to what kind of data I wanted in my science & technology YouTube video title corpus. There was a lot of data that was unusable just because it did not meet the requirements I had set.

First, I had to remove any video titles that were not in English. The language-detection Python module [4] is a direct port from Google's language-detection library. I used this module to accurately detect which video titles were in English. I removed any non-English titles. Some video titles also used symbols such as emojis and punctuation marks. These symbols were also removed from the titles so only alphabetic characters remained. Lastly, I removed any mention of the channel name in the title. I did not feel as if this piece of information was good for the analysis because some video titles spam their channel name and that might have skew the data.

The metadata also needed to be cleaned. Since this information was directly scraped from the YouTube website, values such as views and subscribers were in a string format. I cleaned up the views column by removing any non-numeric characters, appending all the digits together sequentially, and type casting the string to an integer. The subscriber data was a little trickier because its format was abbreviated. I created a small function that looked at the abbreviation and properly converted the string into an integer.

Lastly, I removed any outliers from my data. The outliers in the data came from the videos with a large number of views. To remove the outliers, I generated a five-number summary from my views data and calculated the upper and lower fences [5]. I then cordoned off any data that was either over or below the upper and lower fences respectively.

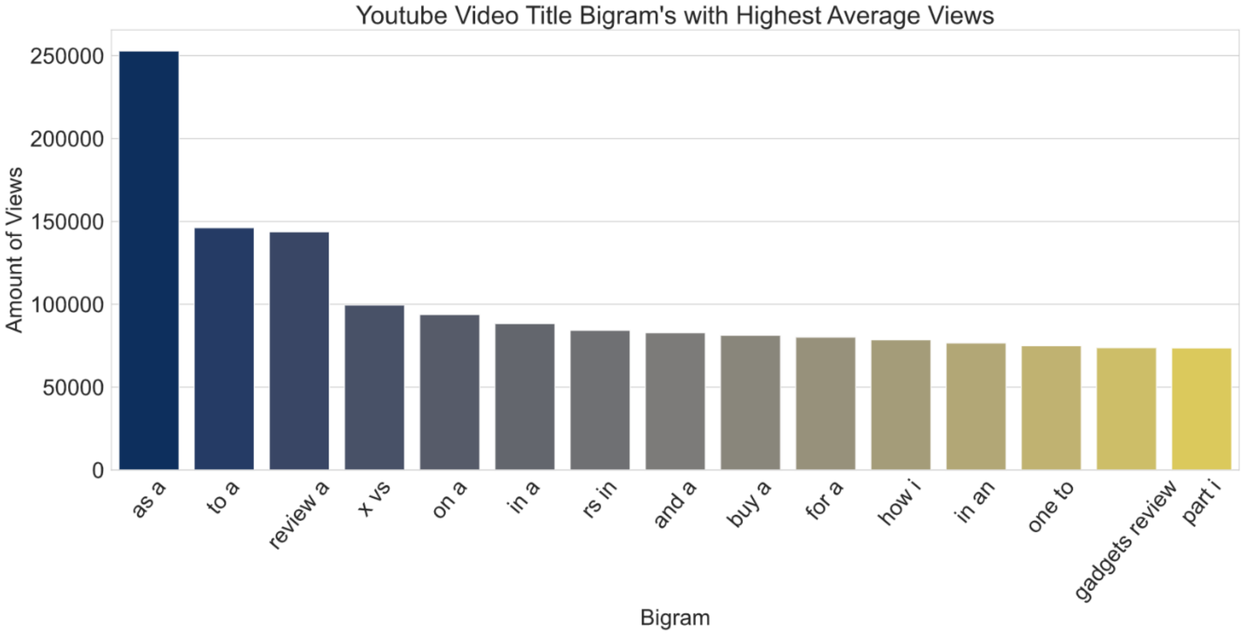
It is important to mention why all the data cleaning was done after the collection. It could be argued that I could have removed any non-English channel names before collecting the data however, most of the non-English titles were in Hindi. Furthermore, these channels typically had an English channel name so there would have been no way of knowing if the title of the video was in English based on the channel name.

IV. Analyzing the Data

The analysis takes a deeper look as to how the language in the title correlates to views. The three main parts of my analysis were word frequency, title length, and the title readability.

A. Word Frequencies

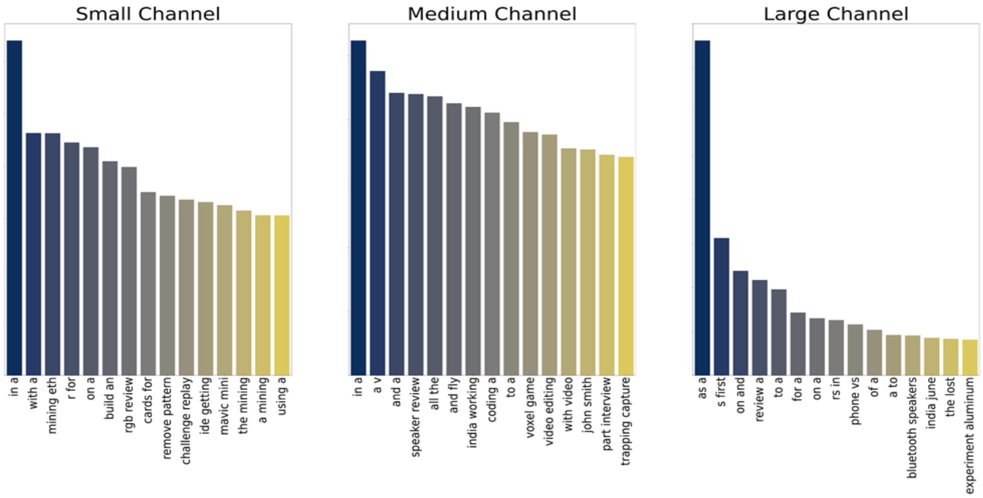
To analyze what kind of language was most frequently used, I looked at the most frequent bigrams from the data. The most frequent bigram turned out to be ”how to” showing up in 419 different video titles. Now, I wanted to see what kind of bigrams were used in more successful videos. To do this, for every bigram, I grabbed the total amount of views associated with the bigram and divided it by the frequency of the bigram. This would give me an average as to which bigrams yielded the most views. The bar chart below displays my findings.



The results were not too surprising as mostly stop words featured in the top 15 list. I did find interesting that the bigram “gadgets review” made the top 15. I would assume there is a large market of gadget reviewers on YouTube so that would explain the high number of views. Below is a concordance of the bigram “gadgets review.”

gaming accessories for pubg mobile ***gadgets review***
realme cheapest gaming phone under ***gadgets review***
realme pro india s first smartphone ***gadgets review***
after one month pubg mobile dknm ***gadgets review***

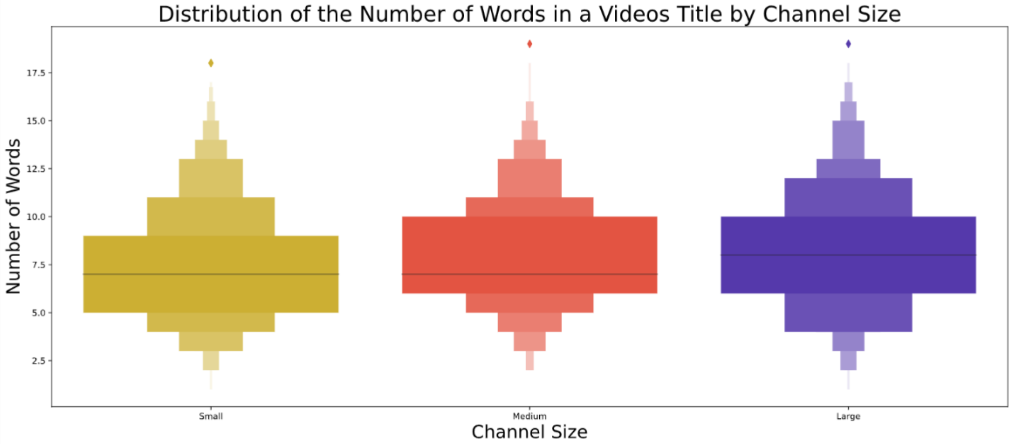
There seems to be a variety of gadget review videos which explains its popularity. To further extend this analysis, I wanted to know if the size of a channel had an impact on the bigrams. I split the data into three equal parts by subscriber count. The data now was equally split based on small, medium, and large sized YouTube channels. I performed the same analysis of the bigrams with the highest on average view count on each of these three splits. The analysis is displayed in the bar charts below.



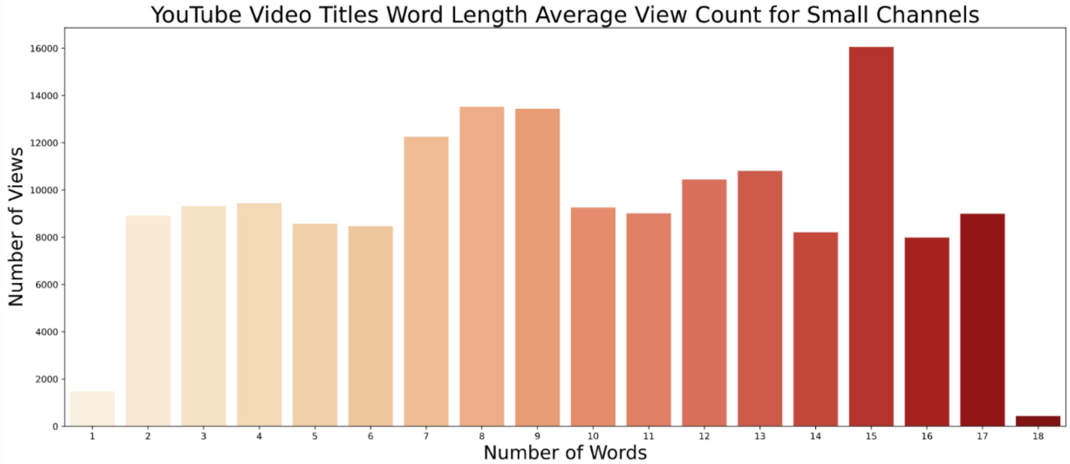
Splitting up the graphs provides more insight as to the most viewed content by different sized YouTube channels. The smaller channels seem to mostly focus on Bitcoin mining. Bitcoin mining was very popular on YouTube a few years back when Bitcoin blew up but seems to not be as popular anymore. There seems to still be a small community which might explain why they dominate the views from the smaller channels. The medium and large channels have no single trend but rather align more closely to the bar chart which displays the title bigrams with highest average views. There is a couple of interesting bigrams from the medium and large channels such as "video editing" and "coding a." These trends show what most viewers are looking for. There seems to be a large group of people trying to learn how to code or edit videos. The larger channels seem to be more diverse in their content.

B. Title Length

Another interesting point of analysis is the word count for the title. Different title lengths might influence the number of views a video gets. A snappier title might bring more users to click a video compared to a longer duller title. I also kept the channel sizes separate to see the differences between different subscriber counts.



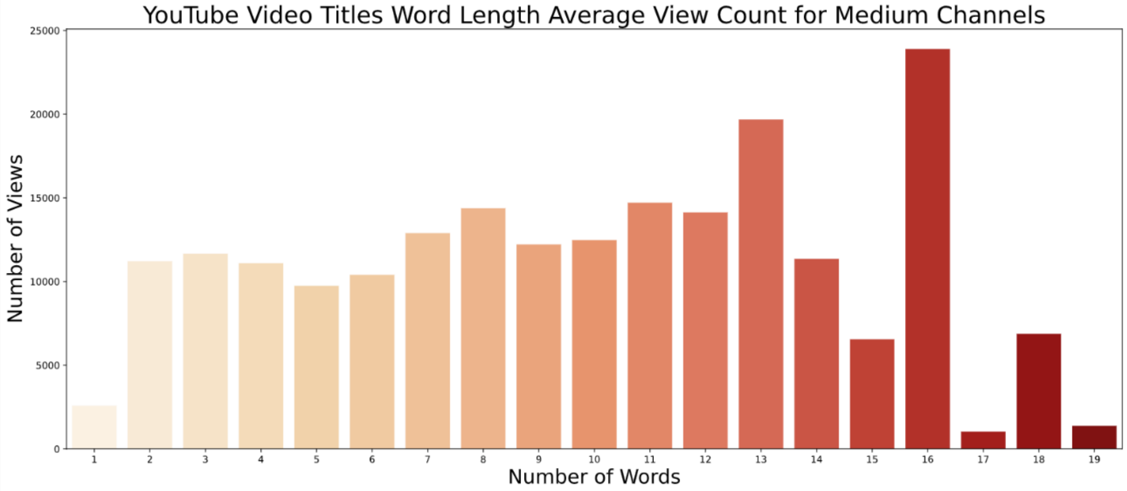
As seen in the graph, the distribution is similar in all three channel sizes. However, the mean of words used in the title increases as the channel size does too. Large channels have the largest distribution and on average, their title word count is around 8 words, while smaller channels on average use around 7 words in their title. As we can see there is a small difference in the title word count for different channel sizes but on average the videos use 7 to 8 words in their title regardless of channel size. Now how does this translate over to views? Let us first look at which word lengths receive the most views on average for small channels.

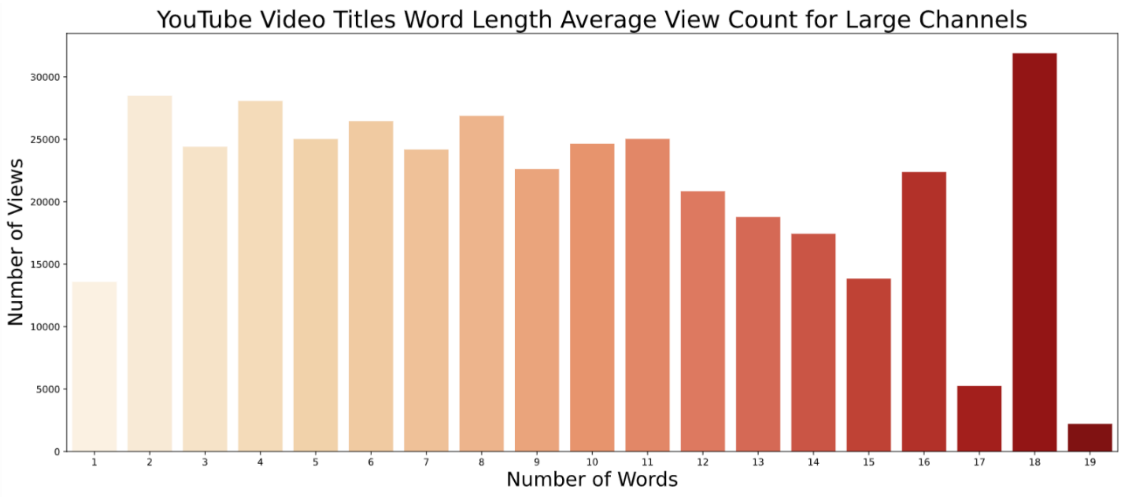


It looks like anywhere from 7 to 9 words in the title yields the greatest number of views. What is surprising here is the sudden spike when looking at 15 words in a title. Having 15 words in your title seems to be a sure way to increase the number of views on a video. A trend I did notice from these videos with 15 words in the title is that they tended to be very specific. Let us take a look at a few:

xiaomi linux android in one duo os bit any pc install on acer tab concept
ubex trading desk tutorial part how to top up you balance and launch a campaign
david schwartz on xrp ledger plus madigan on liquidity voisine on brd wallet ripple drop
get the best footage from your dji drone tips tricks for cinematic footage with filmora

These video titles are very specific and more focused on explaining or teaching some sort of concept. On the other hand, shorter video titles had no specific theme. This is of course because there is a larger sample of shorter video titles. This trend continues with medium sized and large sized channels. Let us look at the results from those.

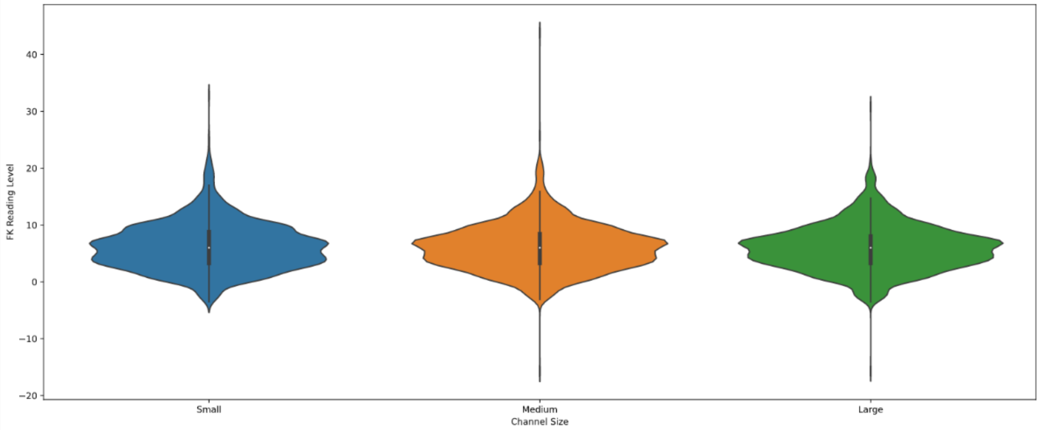




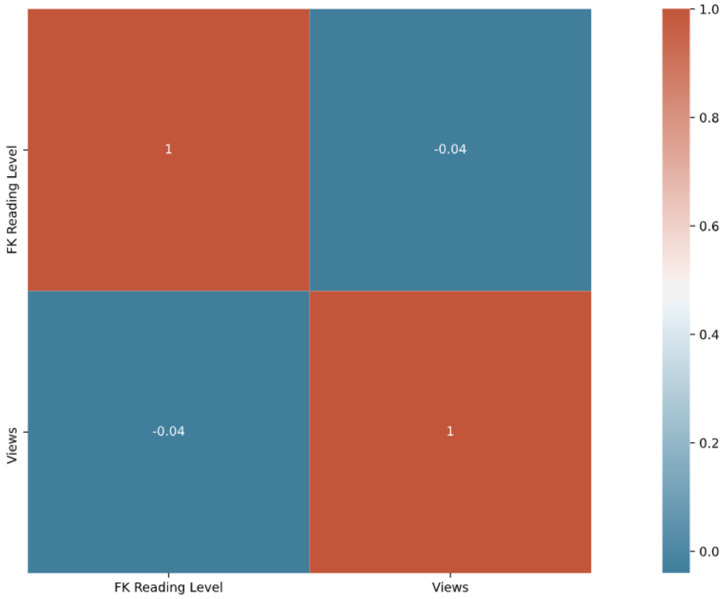
The same trend continues, where longer titles videos seem to give the best results in terms of views. For the medium channels 16 gave the best results as it spiked compared to other word counts. For the larger channels 18 was the peak. An important note is that large channels seem to do better in a larger range of the title word count. Anywhere from 2 to 11 words in the title would help the video achieve a good number of views. On the other hand, the range for small and medium channels is much smaller. For smaller channel, the best range seems to be around anywhere from 7 to 8 words while medium channels have a little more leeway, as their best range is from 7 to 13. The reason for this trend might be that the title of a video is more important to smaller channels since they do not have as strong of a backing as medium or large channels. As a channel grows, they secure a stronger base that will click on their video regardless of title length.

C. Readability

The last point of analysis will be measuring the readability of a title. For this, we use the Flesch Kincaid Readability Tests. To run the test, I used the Textstat Python module [6] which allows for the calculation of statistical features from texts. The library offers various algorithms to determine readability. To calculate the readability, every video title had its readability level calculated. Since the channels were separated based on size, we can see our results below.



As we can see, the distribution of the reading level is similar for all the channel sizes. While some distributions have a larger range, the bulk of the data seems to have the same distribution. The mean of all three channel sizes were 6. What this means Is that this reading level is suitable for a 6th grader. Due to the typical short length of a title, it makes sense that these titles are easier to read. From our previous analysis we also saw that on average most titles are around 7-9 words long and include many common words.



Now let us look at how the readability score effects the views. The graph above shows a heatmap of the correlation between two features. In this case, when we look at the correlation between the readability level and the views, there seems to be no strong linear correlation. This implies that the readability level of the title has directly no impact on the number of views a video will get. This is a bit surprising as I would have assumed that an easier to read title might attract more viewers, but the data does not back that claim up.

V. Conclusion

Trying to understand how the language directly effects the number of views on a video is no easy task. In this report, I analyzed the language by providing insight to the frequent bigrams, title length, and readability score of science & technology YouTube video titles. I then took this language analysis and tried to understand how these language decisions directly effected the number of views a video received. The data used was only a sample of all the science & technology videos on YouTube and with much more data a stronger correlation between language in the video title and views might be achieved.

VI. Looking at the code

Throughout the report I hyperlinked any scripts that I developed for direct access. In this section I will link all the code used. The code is properly commented and documented. Any references for the code are also documented in the code directly and in the code references section.

Collecting Data:

- Scraping Channel Data
- Scraping Video Data

Cleaning & Analyzing Data:

- Cleaning & Analyzing Data

Github Repository:

- YouTube Video Title Language Analysis

References

[1] https://blog.youtube/news-and-events/youtube-at-15-my-personal-journey
[2] https://www.kaggle.com/babikov/youtube-channels-100000
[3] https://github.com/codehax41/Scraping-Youtube-Video-using-Selenium/blob/master/Scrap_Youtube_Description.ipynb
[4] https://pypi.org/project/langdetect/
[5] https://online.stat.psu.edu/stat200/lesson/3/3.2
[6] https://pypi.org/project/textstat/

Code References

[7] https://stackoverflow.com/questions/53775386/how-to-convert-number-mixed-with-character-to-integer-in-pandas
[8] https://www.analyticssteps.com/blogs/extracting-pre-processing-youtube-comments
[9] https://stackoverflow.com/questions/29110950/python-concordance-command-in-nltk
[10] https://stackoverflow.com/questions/48204780/how-to-plot-multiple-figures-in-a-row-using-seaborn
[11] https://seaborn.pydata.org/examples/grouped_violinplots.html
[12] https://seaborn.pydata.org/examples/many_pairwise_correlations.html