BT4221: Big Data Techniques and Technologies

Group Report

Music
SUBSCRIBE
Home
Videos
Playlists
Channels
Top Music Tracks
We Are One (Ole Ola) [The Official 2014 FIFA World Cup Song] (Olodum Mi
PitbullVEVO
vevo
4:06
Demi Lovato - Let It Go (from "Frozen") [Official]
DemiLovatoVEVO
3:51
Katy Perry - Dark Horse (Official) ft. Juic
KatyPerryVEVO
3:46
More
Top R&B Tracks

Going Big (Data) on YouTube

Prepared by: Group 1

Chua Kai Bing
A0185606Y
kai_bing.chua@u.nus.edu

Goh Jia Yi
A0185610J
gohjiayi@u.nus.edu

Ngoc Linh Chi Nguyen
A0170767W
nguyenngoclinhchi@u.nus.edu

Tan Zen Wei
A0188424X
zenwei.tan@u.nus.edu

1. Introduction

1.1 Business Case

Living in the 4.0 technology era, YouTube has inevitably become a platform that many uses on a day-to-day basis. According to Aslam (2020), YouTube has over 2 billion monthly users who watch millions of hours of content daily. As of mid-June 2018, there have been over 5 billion videos shared on the platform by over 50 million users.

One main goal that many YouTubers have is to monetise their videos and make a profit from them. The more views and the longer watch time one video has, the higher the profit generated. However, with millions of people uploading videos on YouTube every day, it is hard to attract viewers with intense competition (Berman, 2020).

Fortunately, with significant improvement in big data technologies, analysing large amounts of data has become possible. By making use of these available technologies, insights can be generated from channel and video data collected from popular YouTubers; these insights can help small YouTubers to identify possible ways to grow their channel and increase YouTube views.

1.2 Motivation

By generating insights on the most significant factors that can help to increase YouTube video views, this project aims to help YouTubers grow their channels.

1.3 Hypotheses

In this project, the following hypotheses will be tested.

A. "A positive sentiment title in a YouTube video will affect views"
B. "Having humans in the thumbnail images of a YouTube video will affect views"

2. Data Collection

The YouTube Data API v3 by Google Developers was utilised to collect YouTube channel and video data. This API provides public information available on the platform, such as the channels, videos, comments and ratings.

For this project, the scope will be focused on a subset of popular YouTube channels. The team first gathered a list of the Top 10 most subscribed English-speaking channels across 17 categories as defined by Social Blade—a platform that tracks social media statistics and analytics across platforms such as YouTube, Twitch, Instagram, and Twitter. The 17 categories are ‘Auto & Vehicles’, ‘Comedy’, ‘Education’, ‘Entertainment’, ‘Film’, ‘Gaming’, ‘How to & Style’, ‘Made for Kids’, ‘Music’, ‘News & Politics’, ‘Nonprofit & Activism’, ‘People & Blogs’, ‘Pets & Animals’, ‘Science & Technology’, ‘Shows’, ‘Sports’ and ‘Travel’. The ‘Shows’ category was eventually removed as the low subscriber and video counts for the channels within this category, provided us with little video data to analyse. Thus, the project will be focusing on a total of 160 YouTube channels across 16 categories. Information about these channels such as their channel ID, name and category was found from Social Blade and manually keyed into a .csv file.

Based on the channel IDs gathered, the team used three channel resources provided by the YouTube API — snippet, statistics and content details, to gather the following channel data.

- **Snippet** provides information about a channel such as the channel name and description, creation time and region.
- **Statistics** contains the performance metrics such as the view, subscriber and video count of a channel.
- **Content Details** provides the playlist ID, which was used to locate the video collections of the channel.

Thereafter, using the playlist ID of each channel, the snippet and statistics video resources from the API were utilised to collect the following video data.

- **Snippet** contains information such as the video title, description and URL of the thumbnail images.
- **Statistics** provides the performance metrics such as the like, dislike, comment and view count of the video.

Since the YouTube API restricts the number of videos that can be retrieved from each channel, only the most recent 20,000 videos before 2015 for each channel were extracted. The time filter was put in place to ensure that the analysis will be relevant in today’s context.

Besides the above limitation, each Google Developer project provides one API key that was limited to 10,000 queries per day. To work around this limit, the team had to split up the 16 categories and run the code for data extraction on separate occasions with different API keys before joining the full data together to achieve efficiency.

A total of 408,690 videos from 160 channels were extracted at this stage. The full description of our dataset can be found in Appendix B: Data Description.

3. Data Cleaning and Preprocessing

Upon manual inspection of the data collected, the ‘favouriteCount’, ‘viewCount’, ‘likeCount’, ‘dislikeCount’, and ‘commentCount’ fields in the video dataset contained a significant number of missing values. This phenomenon is observed when YouTubers chose to hide these statistics from the public.

Having null values in the dataset can prevent meaningful statistical analysis from being carried out. The following methods were adopted to tackle this issue.

3.1 Removing Missing Values

The ‘favouriteCount’ field was removed since the entire column was made up of null values. Also, the rows with missing ‘likeCount’, ‘dislikeCount’ and ‘viewCount’ were dropped as they only accounted for a small portion of the overall dataset (roughly 100 rows).

3.2 Imputing Missing Values

Rows with missing values for ‘commentCount’ were retained since they made up a significant portion of the data (roughly 30,000 rows) and the team decided to impute the missing values.

To fill the missing values in ‘commentCount’, the team attempted several methods to obtain values that can be representative of the data. The video data was split into training and testing data with a ratio of 8:2. Subsequently, the mean squared error (MSE) and R-squared values were computed after replacing the missing values with values obtained from each method.

Initially, the team tried using mean and zeros to impute the missing ‘commentCount’ values. However, this resulted in high MSE and low R-squared scores—an indication that the replaced values were not representative of the actual ones. Hence, linear regression was used instead to fill in these missing values. For each of the 16 categories, a linear regression model was built and applied to the rows with missing ‘commentCount’ values. The independent variables used in the model were ‘viewCount’, ‘likeCount’ and ‘dislikeCount’. These predicted values gave better MSE and R-squared values for all categories as compared to the previous method, with the exception of the ‘Made for Kids’ category which showed better metrics when the mean values were used. Thus, the mean ‘commentCount’ value was used to fill the missing ‘commentCount’ values in the ‘Made for Kids’ category, while the rest of the categories used the predicted values from their respective linear regression models.

This decreased the size of our dataset to 407,093 rows for feature extraction.

4. Feature Extraction

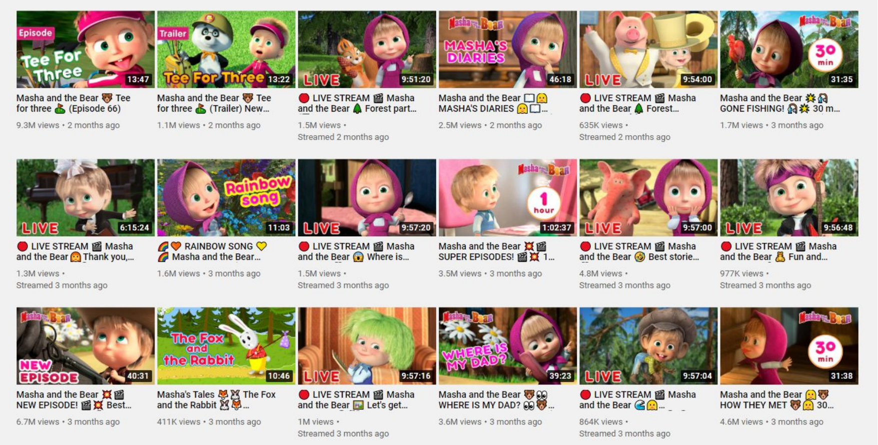
Feature Extraction is needed as the data collected may not be sufficient to build a good machine learning model most of the time. The two most important variables that can affect views of a YouTube video is the title and thumbnail image as they play an essential role in capturing the audience's attention. By optimising these two variables, it can potentially help to increase the chance of YouTube users viewing a specific video.

Additional features that can account for differences between the title and thumbnail images of the videos would be insightful for this project. Thus, in this section, sentiment analysis and part-of-speech tagging were applied to the video titles to understand the sentiment and sentence structure of the title better. Object detection was also applied to the video thumbnail images to identify the number of people present in it.

4.1 Sentiment Analysis

Sentiment analysis, or opinion mining, is the field of study related to analysing opinions, sentiments, evaluations, attitudes, and emotions of users that are expressed on social media and other online resources. The evolution of social media sites has also attracted users to express their opinions or sentiments about the videos they watch on video sharing sites, such as YouTube.

Out of all the video titles in the dataset, 3,264 of the titles contain emojis in them. 'Masha and The Bear' is a channel from the Film category that frequently incorporates emojis in their titles, as seen in Figure 1 below.



Masha and the Bear 🐻 Tee for three ⛳ (Episode 66)
9.3M views • 2 months ago

Masha and the Bear 🐻 Tee for three ⛳ (Trailer) New...
1.1M views • 2 months ago

🔴 LIVE STREAM 🎬 Masha and the Bear 🌲 Forest part...
1.5M views •
Streamed 2 months ago

Masha and the Bear 📖🙂 MASHA'S DIARIES 🙂📖...
2.5M views • 2 months ago

🔴 LIVE STREAM 🎬 Masha and the Bear 🌲 Forest...
635K views •
Streamed 2 months ago

Masha and the Bear 💥🎧 GONE FISHING! 🎧💥 30 m...
1.7M views • 3 months ago

🔴 LIVE STREAM 🎬 Masha and the Bear 🙂 Thank you,...
1.3M views •
Streamed 3 months ago

🌈🧡 RAINBOW SONG 💛 🌈 Masha and the Bear...
1.6M views • 3 months ago

🔴 LIVE STREAM 🎬 Masha and the Bear 😱 Where is...
1.5M views •
Streamed 3 months ago

Masha and the Bear 💥🎬 SUPER EPISODES! 🎬💥 1...
3.5M views • 3 months ago

🔴 LIVE STREAM 🎬 Masha and the Bear 🤣 Best storie...
4.8M views •
Streamed 3 months ago

🔴 LIVE STREAM 🎬 Masha and the Bear 🧸 Fun and...
977K views •
Streamed 3 months ago

Masha and the Bear 💥🎬 NEW EPISODE! 🎬💥 Best...
6.7M views • 3 months ago

Masha's Tales 🦊🐰 The Fox and the Rabbit 🐰🦊...
411K views • 3 months ago

🔴 LIVE STREAM 🎬 Masha and the Bear 🖼️ Let's get...
1M views •
Streamed 3 months ago

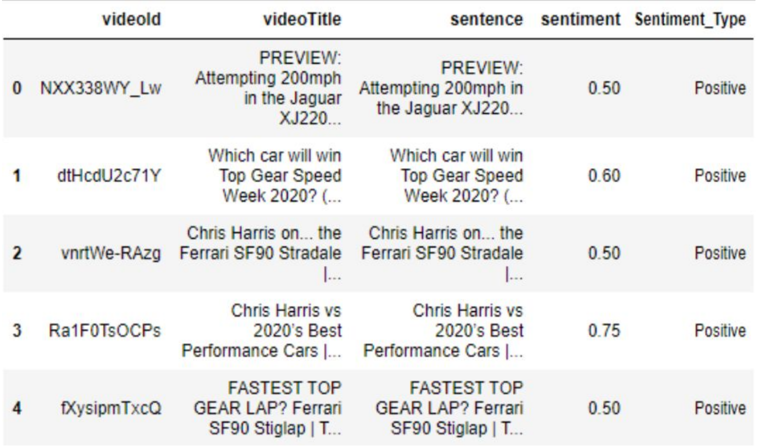
Masha and the Bear 🐻👀 WHERE IS MY DAD? 👀🐻...
3.6M views • 3 months ago

🔴 LIVE STREAM 🎬 Masha and the Bear 🌊🙂...
864K views •
Streamed 3 months ago

Masha and the Bear 🙂🐻 HOW THEY MET 🐻🙂 30...
4.6M views • 3 months ago

Figure 1: Video page of 'Masha and The Bear' Channel

Just like how emojis can be used to replace words within messages, they can also help to add extra meaning and emotion into video titles when used, by conveying the tone and non-verbal context behind the sentences. Hence, all emojis in our video titles were decoded into words with the use of the *emoji* Python library. For example, the emoji "💃" becomes "woman dancing" after decoding.



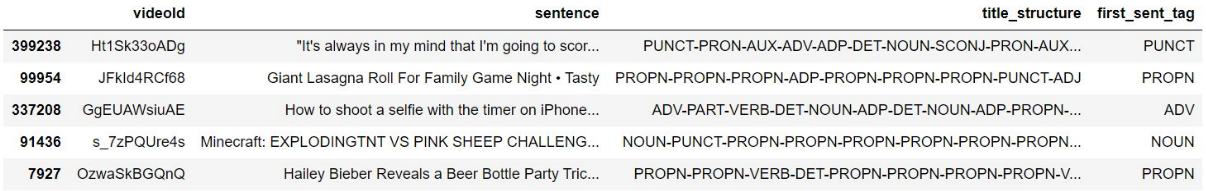
	videoId	videoTitle	sentence	sentiment	Sentiment_Type
0	NXX338WY_Lw	PREVIEW: Attempting 200mph in the Jaguar XJ220...	PREVIEW: Attempting 200mph in the Jaguar XJ220...	0.50	Positive
1	dtHcdU2c71Y	Which car will win Top Gear Speed Week 2020? (...	Which car will win Top Gear Speed Week 2020? (...	0.60	Positive
2	vnrtWe-RAzg	Chris Harris on... the Ferrari SF90 Stradale |...	Chris Harris on... the Ferrari SF90 Stradale |...	0.50	Positive
3	Ra1F0TsOCPs	Chris Harris vs 2020's Best Performance Cars |...	Chris Harris vs 2020's Best Performance Cars |...	0.75	Positive
4	fXysipmTxcQ	FASTEST TOP GEAR LAP? Ferrari SF90 Stiglap | T...	FASTEST TOP GEAR LAP? Ferrari SF90 Stiglap | T...	0.50	Positive

Figure 2: Results from Sentiment Analysis

Afterwhich, sentiment analysis was conducted on the video titles using the *TextBlob* Python library, and the results can be seen in Figure 2 above. Sentiment refers to the sentiment score of the sentence, which lies in the range of [-1, 1]. The higher the value, the more positive the sentence. The titles were then categorised into 3 different sentiment types—'Positive' (sentiment > 0), 'Neutral' (sentiment = 0) and 'Negative' (sentiment < 0).

4.2 Part-of-Speech (PoS) Tagging

Part-of-Speech (PoS) Tagging is the process of labelling a word in a corpus to a corresponding part of a speech tag, based on its context and definition. YouTube videos are known for their clickbait titles. By understanding the sentence structure of these video titles, it may deliver additional insights on whether it makes an impact on the view count. The sentence structure of a sentence can be obtained by applying the Coarse-grained PoS tags for video titles. For this step, the *spaCy* Python library was used.



	videoId	sentence	title_structure	first_sent_tag
399238	Ht1Sk33oADg	"It's always in my mind that I'm going to scor...	PUNCT-PRON-AUX-ADV-ADP-DET-NOUN-SCONJ-PRON-AUX...	PUNCT
99954	JFkId4RCf68	Giant Lasagna Roll For Family Game Night • Tasty	PROPN-PROPN-PROPN-ADP-PROPN-PROPN-PROPN-PUNCT-ADJ	PROPN
337208	GgEUAWsiuAE	How to shoot a selfie with the timer on iPhone...	ADV-PART-VERB-DET-NOUN-ADP-DET-NOUN-ADP-PROPN-...	ADV
91436	s_7zPQUre4s	Minecraft: EXPLODINGTNT VS PINK SHEEP CHALLENG...	NOUN-PUNCT-PROPN-PROPN-PROPN-PROPN-PROPN-PROPN...	NOUN
7927	OzwaSkBGQnQ	Hailey Bieber Reveals a Beer Bottle Party Tric...	PROPN-PROPN-VERB-DET-PROPN-PROPN-PROPN-PROPN-V...	PROPN

Figure 3: Results from PoS Tagging Model

Figure 3 above shows the results from the PoS Tagging model. The results from our model application showed us more than 300,000 different combinations of sentence structures. Hence, the team has decided to narrow down our scope to focus on the PoS for the first word of the video title, as they play a more important role in attracting viewer's attention, given that titles are truncated when viewed on small screen devices such as smartphones. After this feature extraction step, additional 17 features for the following PoS tags was created - 'PROPN', 'DET', 'ADJ', 'VERB', 'NOUN', 'NUM', 'AUX', 'PUNCT', 'ADV', 'ADP', 'X', 'PRON', 'SYM', 'SCONJ', 'INTJ', 'PART', 'CCONJ' (Refer to Appendix C: Part-of-Speech (PoS) Tags for more details).

4.3 Object Detection

Although thumbnail images are relatively crucial in attracting a user to view a video, not much information was given about them in the YouTube data extracted, except for the Uniform Resource Locator (URL) of the thumbnail images. For the team to effectively evaluate the effects that thumbnail images have on view count, object detection can be performed to generate additional insights.

Out of all the video URLs in the dataset, 65 thumbnail images could not be retrieved. This could have happened if the content owner changed the privacy settings of the video between the time when the YouTube video data was extracted and the time when the thumbnail images were downloaded.

Due to the size of the dataset, the time taken to perform object detection inference on all thumbnail images was extremely slow without a Graphics Processing Unit (GPU). Fortunately, given access to Amazon Web Service (AWS) S3 and SageMaker, the processing speed was improved significantly. The object detection model, GluonCV You Only Look Once (YOLO) v3, was deployed. Given an input image, the model is able to return the object coordinates and category predictions, as seen in Figure 4 below.

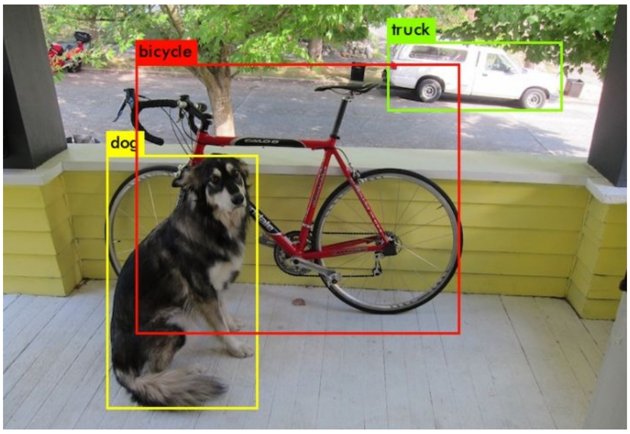
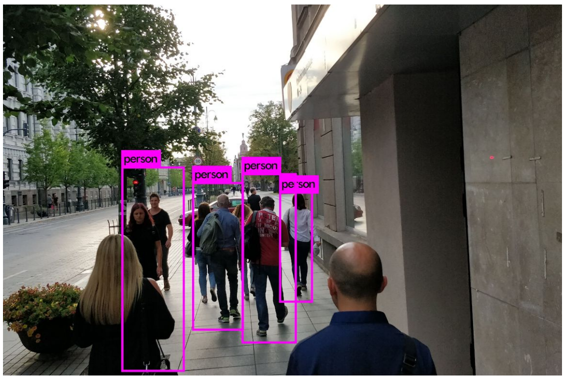


Figure 4: Example Output from Object Detection Model

YOLO was trained on the COCO dataset with 80 common object categories (refer to Appendix D: Object Detection Model). As Hypothesis B of the project was to test if the presence of humans has an effect of view count, the model was tailored only to detect for the presence of people. For every thumbnail image, a numerical value of the number of people detected ('humanCount') and a binary value indicating if people were present ('humanPresent') was stored. The threshold of the model was set to 0.2, meaning only objects predicted to be .



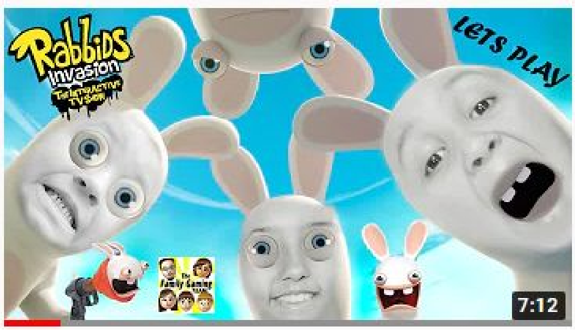


Figure 5: Object Detection Model Misclassification (left)
Figure 6: Animated Thumbnail image on YouTube (right)

Despite YOLO being a state-of-the-art object detector, there are still some caveats of the model. Firstly, misclassification may occur when a person present is undetected or when another object is detected as a person, as seen in Figure 5 above. Such occurrences are common when there is occlusion[1], which is an area for improvement for most computer vision models. Secondly, a large number of YouTube thumbnail images contained animations, with some being fully animated while others were a mix of photos and animation. When the model was applied to animated thumbnail images such as one seen in Figure 6 above, it was unable to identify the presence of people accurately. This is because animated images were not present in the COCO dataset that was used to train the model. With this, the caveats of the object detection model should be kept in mind when performing our analysis.

4.4 Others

Besides the feature extraction methods mentioned above that involved the use of machine learning models, some basic methods were used as well.

First, the number of words in each video title was tabulated and stored as a new feature - 'titleLen'. Next, the average view count of each channel was calculated by taking the total view count divided by the total number of videos for each specific channel. This additional feature was stored as 'avgViewCount'. Lastly, the published timing of videos from the feature 'publishedAt' was further divided into three categories—the time of day, day of week and season.

- **Time of Day:** 6 categories—'Early Morning', 'Morning', 'Afternoon', 'Evening', 'Night' or 'Late Night'.
- **Day of Week:** 7 categories—'Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday', 'Saturday' or 'Sunday'.
- **Season:** 4 categories—'Spring', 'Summer', 'Autumn' or 'Winter'.

The size of our dataset decreased to 405,759 rows after feature extraction was conducted.

[1] Occlusion occurs when two or more objects come too close and seemingly merge or combine with each other, causing the computer vision to be unable to identify or track objects.

5. Data Exploration

In this section, visualisations were generated to gain further insights on our dataset.

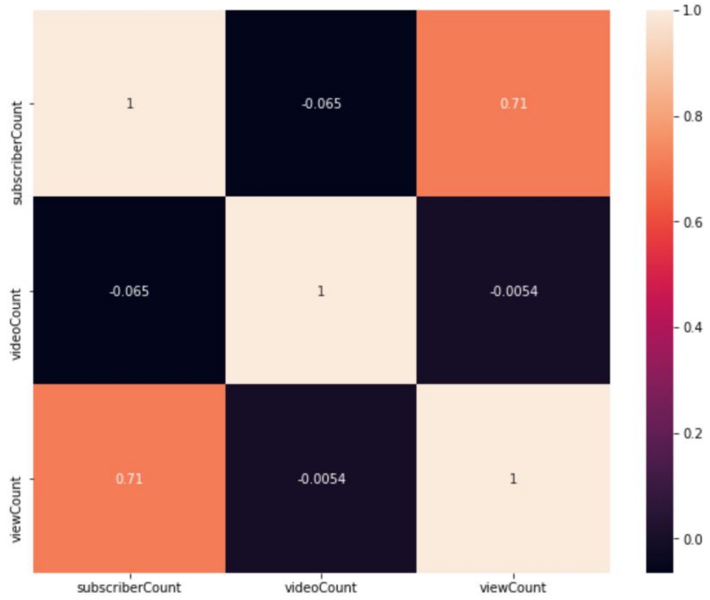


Figure 7: Correlation Matrix for Numerical Fields of Channel Data

Based on the correlation matrix in Figure 7 above, 'subscriberCount' and 'viewCount' have a strong positive correlation of 0.71 while the other fields have a very weak negative correlation with each other.

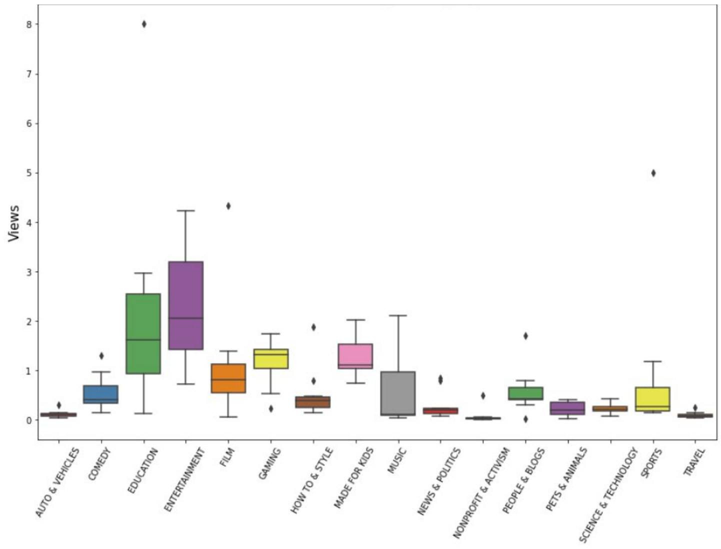


Figure 8: Box Plot of View Count by Category

When comparing the median view counts across categories, 'Non-Profit & Activism' channels had the lowest view counts, while 'Entertainment' channels had the highest, as seen in Figure 8 above.

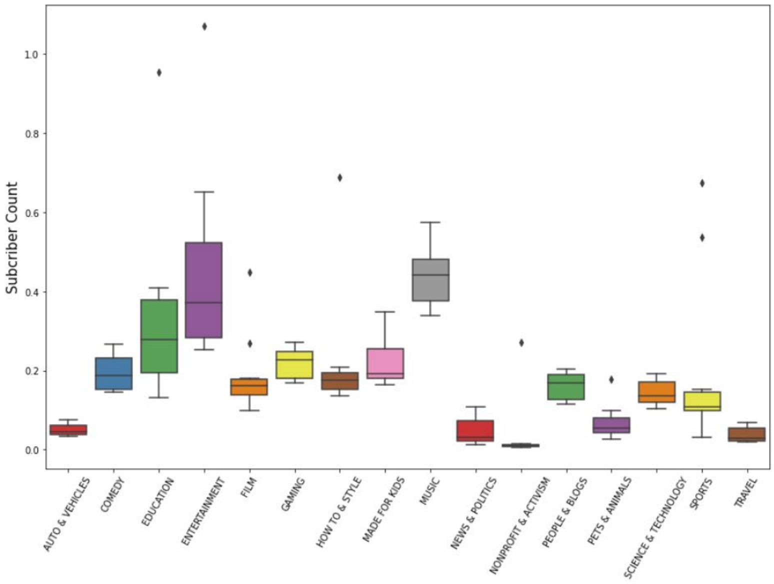


Figure 9: Box Plot of Subscriber Count by Category

Similarly, when observing the median of subscriber counts across categories, ‘Non-Profit & Activism’ channels had the lowest subscriber counts, while ‘Entertainment’ channels had the highest, as seen in Figure 9 above. This phenomenon can be supported by the strong correlation between subscriber and view count, as shown in Figure 7 above.

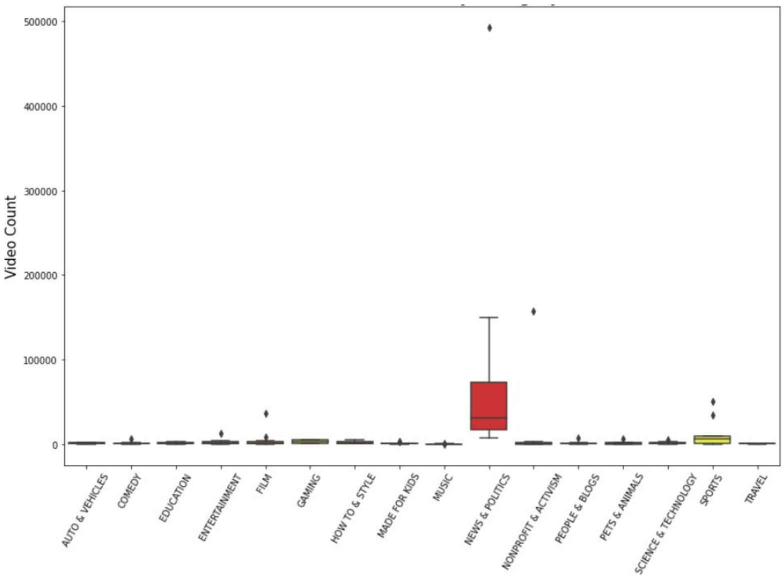


Figure 10: Box Plot of Video Count by Category

When observing the median of video counts across categories, ‘News & Politics’ channels were the highest with a median of approximately 20,000. On the other hand, it is evident from Figure 10 above that channels from other categories have relatively lesser videos. Upon further inspection, around 25% of the ‘News & Politics’ channels have more than 100,000 videos. This finding is not surprising as ‘News & Politics’ channels have to regularly report on the latest news from all over the world daily, resulting in the high volume of videos posted on their channels.

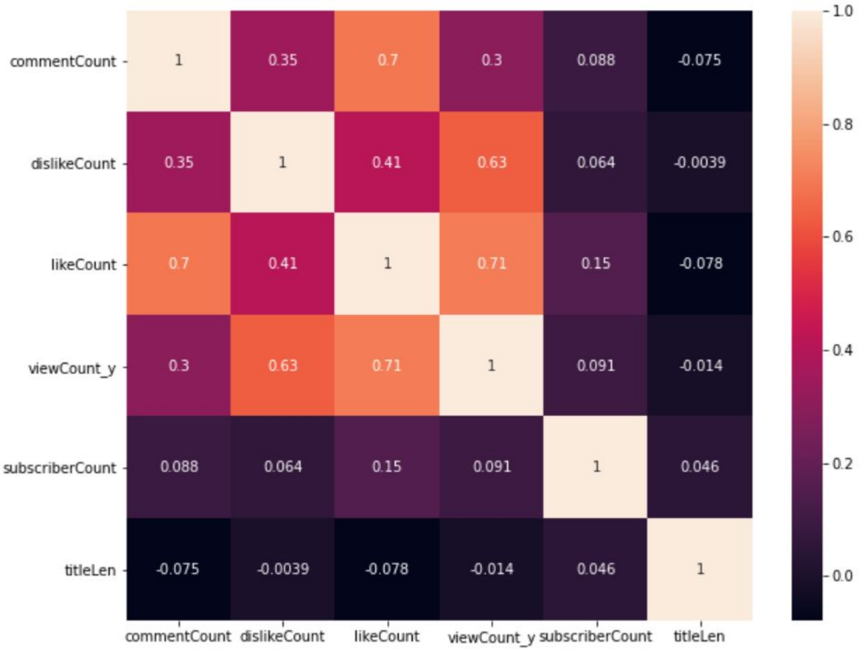


Figure 11: Correlation Matrix for Numerical Fields of Video Data

Figure 11 above shows the correlation matrix for the numerical fields of video data. It is observed that the features ‘likeCount’, ‘dislikeCount’, and ‘commentCount’ are positively correlated to one another. In particular, ‘likeCount’ and ‘dislikeCount’ have a strong positive correlation of 0.71 and 0.63 respectively with ‘viewCount’. It can also be observed that there is a strong positive correlation of 0.7 between ‘likeCount’ and ‘commentCount’.

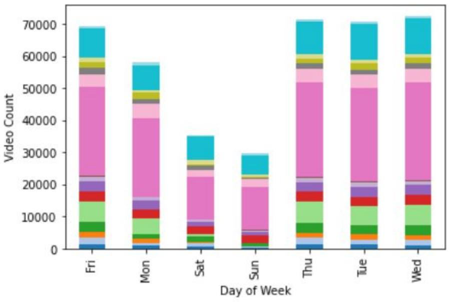


Figure 12: Publish Day of Videos

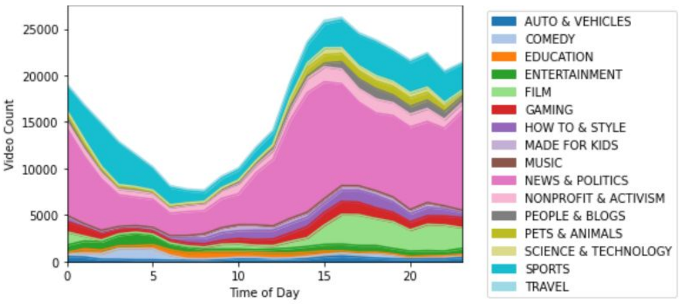


Figure 13: Publish Time of Videos

While many may expect more videos to be posted on weekends since it is likely that more users would be online then, our results say otherwise. As seen in Figure 12 above, weekends have the least number of videos being published. It can also be observed that the peak hours of publishing videos is around 3 to 5 pm, as seen in Figure 13 above. This may be due to the increase in user traffic during this timeframe, hence causing YouTubers to prefer publishing their videos at this timing. In general, the video publishing trend for day and time across categories are similar.

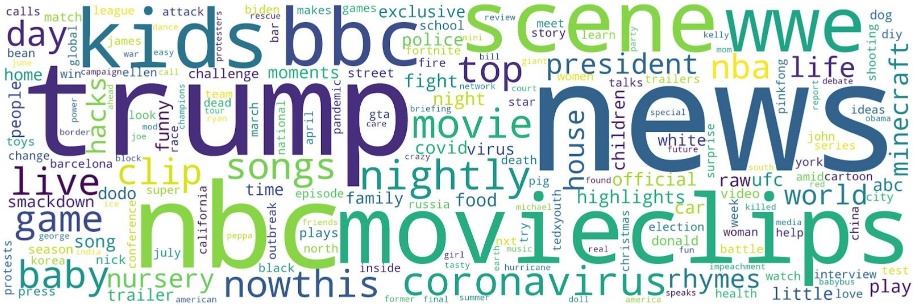


Figure 14: Word Cloud of Truncated Video Titles

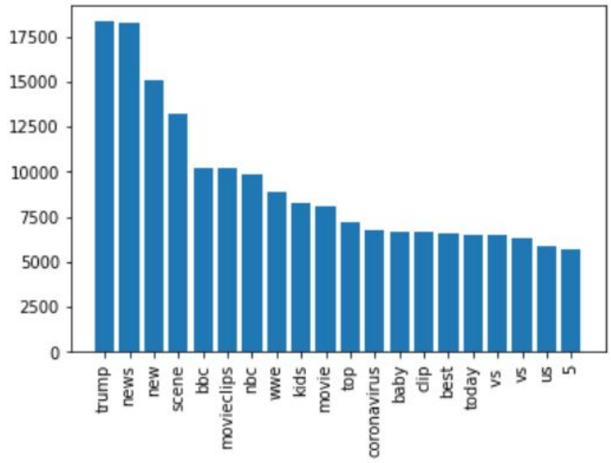


Figure 15: 20 Most Frequent Words in Truncated Video Titles

According to Figure 14 and 15 above, the most frequent words appearing in video titles is ‘news’, followed by ‘trump’. From Figure 10 earlier, it was observed that the ‘News & Politics’ channels had published a relatively greater number of videos as compared to the other categories. The channels in this category may have published many news videos relating to Donald Trump, the previous US president, within the past 4 years of his presidency, thus explaining the high frequency of these 2 words in the dataset.

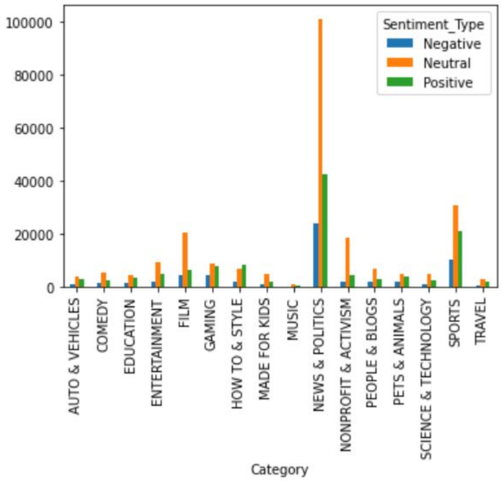


Figure 16: Count of Sentiment Types by Category

Among all the sentiments of the video titles detected by the sentiment analysis model, it can be observed that most of the titles were neutral. At the same time, there was only a small portion of negative titles, as seen in Figure 16 above. There is generally a higher number of positive sentiments compared to negative sentiments. This implies that YouTubers tend to use more positive sentiment titles for their videos.

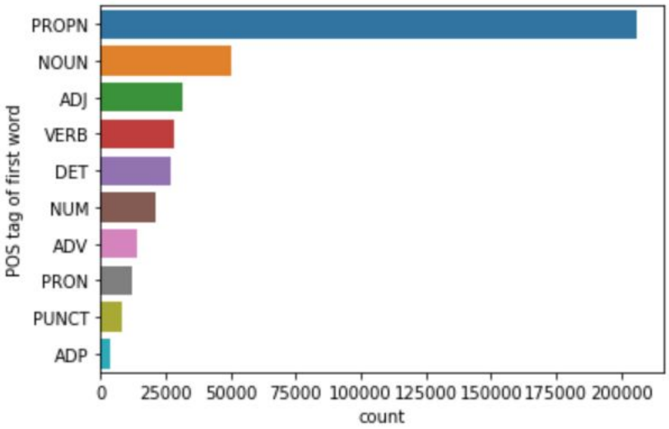


Figure 17: Count Plot of Top 10 PoS Tags Used in the First Word of Video Title

Figure 17 above shows the Top 10 most frequent PoS tags used for the first word of a video title. Approximately 50% of the videos in the dataset uses a proper noun as the first word of their title. A proper noun names a particular person, place, or thing. For example, the title "Las Vegas Hangover Food Taste Test" starts with a place, Las Vegas.

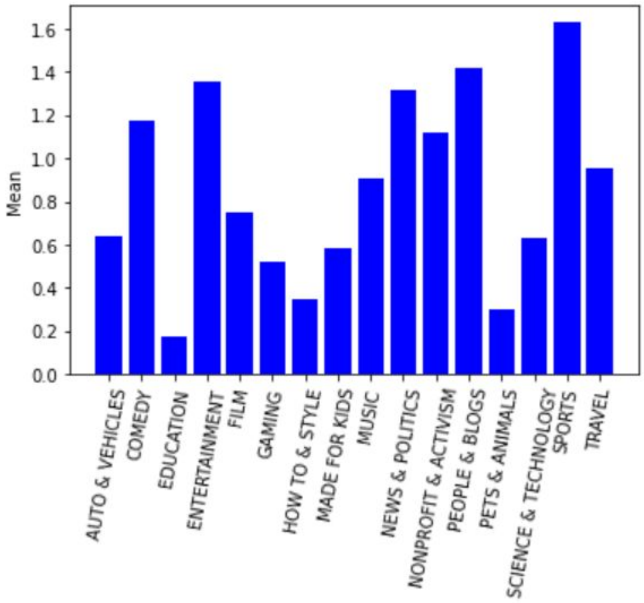


Figure 18: Average Human Count in Video Thumbnail Images by category

Figure 18 above shows the average number of humans present in the thumbnail images of the video for each category. It can be observed that the 'Sports' category has the most number of humans in their thumbnail images, while the 'Education' category has the least. This is not an unexpected finding since the thumbnail images of sports videos usually consist of the players to allow viewers to identify the team that is competing in the video.

6. Feature Selection

For feature selection, only important features that can help to improve the performance of our machine learning models are selected. This is done to cut down training time with the huge dataset.

Table 1 below shows the 58 features (4 numerical, 54 categorical) that were generated from the Feature Extraction stage. One-hot encoding was used to process the categorical features into binary values.

Variable Type	Data Type	Features
Dependent (y)	Numerical	'viewCount'
Independent (X)	Numerical	'titleLen', 'subscriberCount', 'avgViewCount', 'humanCount'
	Categorical	'Friday', 'Monday', 'Saturday', 'Sunday', 'Thursday', 'Tuesday', 'Wednesday', 'Afternoon', 'Early Morning', 'Evening', 'Late Night', 'Morning', 'Night', 'Autumn', 'Spring', 'Summer', 'Winter', 'AUTO & VEHICLES', 'COMEDY', 'EDUCATION', 'ENTERTAINMENT', 'FILM', 'GAMING', 'HOW TO & STYLE', 'MADE FOR KIDS', 'MUSIC', 'NEWS & POLITICS', 'NONPROFIT & ACTIVISM', 'PEOPLE & BLOGS', 'PETS & ANIMALS', 'SCIENCE & TECHNOLOGY', 'SPORTS', 'TRAVEL', 'Negative', 'Neutral', 'Positive', 'titlePROPN', 'titleDET', 'titleADJ', 'titleVERB', 'titleNOUN', 'titleNUM', 'titleAUX', 'titlePUNCT', 'titleADV', 'titleADP', 'titleX', 'titlePRON', 'titleSYM', 'titleSCONJ', 'titleINTJ', 'titlePART', 'titleCCONJ', 'humanPresence'

Table 1: Features before Feature Selection

The ‘likeCount’, ‘dislikeCount’, ‘commentCount’ and ‘viewCount’ of the videos are variables collected at the start of our project, which can be used to measure video engagement levels. However, as the main focus of this project is to see how the features can affect view count, only the feature ‘viewCount’ was used as the dependent variable.

Feature selection was performed in three separate stages—categorical features, numerical features and all features together. Based on the results, features that were deemed important by these methods will be kept for the model building stage while the rest will be dropped.

6.1 Categorical Features - ANOVA

ANOVA test was conducted on the categorical input variables.

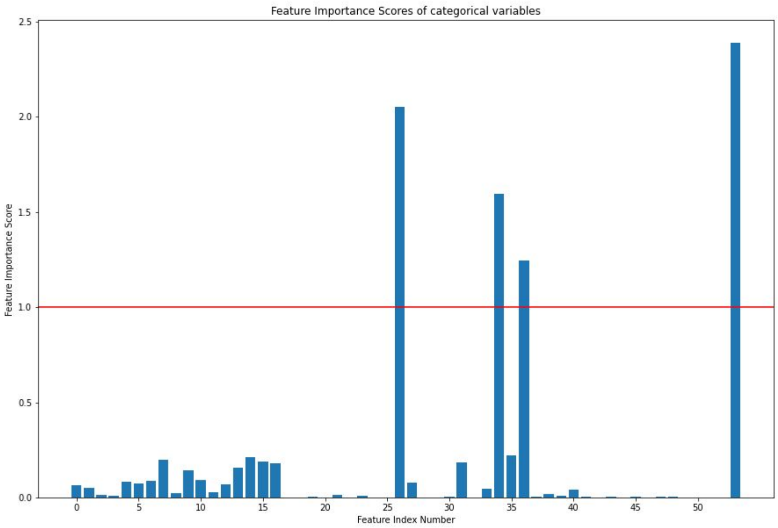


Figure 19: ANOVA Test for Categorical Variables

The scores of the features from the test can be seen in Figure 19 above. The x-axis represents the index of the feature, while the y-axis shows the feature importance score. Features that have a feature importance score of more than 1 were selected, while the remaining features were dropped. Table 2 below shows the selected categorical features from this stage.

Index	Feature	Score
26	HOW TO & STYLE	2.051160
34	SPORTS	1.596341
36	Negative	1.246622
53	titleINTJ	2.386547

Table 2: Selected Categorical Features

6.2 Numerical Features - Pearson’s Correlation

Pearson’s Correlation test was conducted on the output and numerical input variables.

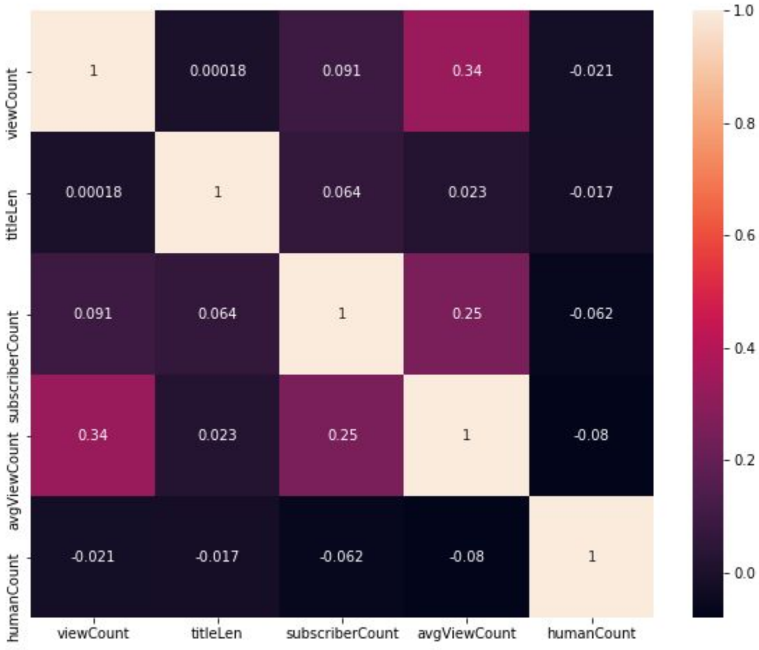


Figure 20: Pearson’s Correlation Test for all Numerical Variables

From the correlation matrix in Figure 20 above, it can be observed that none of the numerical input variables are strongly correlated to one another, or the output variable ‘viewCount’. This implies that the features are not linearly dependent on one another, and hence has no similar effect on the output variable. Hence, none of the numerical features were dropped.

6.3 All Features - ANOVA

ANOVA test was conducted on all input variables.

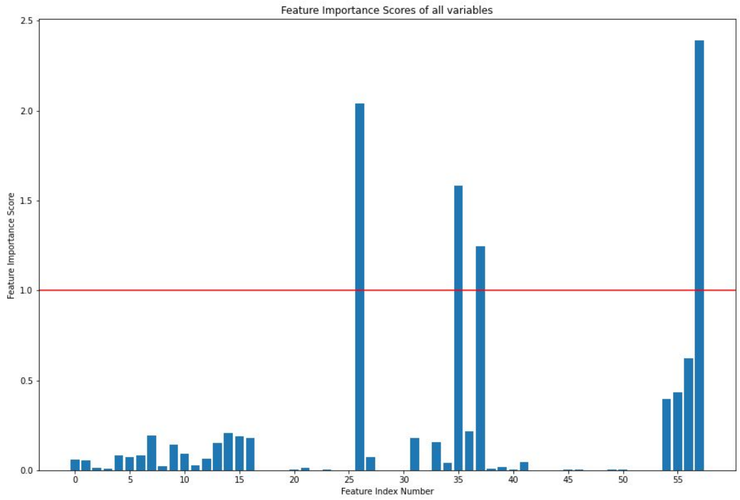


Figure 21: ANOVA Test for All Variables

The scores of the features from the test can be seen in Figure 21 above. Features that have a feature importance score of more than 1 were selected, while the remaining features were dropped. Table 3 below shows the selected features from this stage.

Index	Feature	Score
26	HOW TO & STYLE	2.041418
35	TRAVEL	1.582291
37	Negative	1.244789
57	subscriberCount	2.389420

Table 3: Selected Categorical+Numerical Features

6.4 Selected Features

Taking the union of the results from the above 3 models, Table 4 below shows the features which will be used for model building. 4 numerical and 5 categorical features were selected, resulting in a total of 9 independent variables to be used in our final model.

Variable Type	Data Type	Features
Dependent (y)	Numerical	'viewCount'
Independent (X)	Numerical	'titleLen', 'subscriberCount', 'avgViewCount', 'humanCount'
	Categorical	'HOW TO & STYLE', 'SPORTS', 'TRAVEL', 'Negative', 'titleINTJ'

Table 4: Features after Feature Selection

Out of all the 16 video categories, the 'How to & Style', 'Sports', and 'Travel' category features were selected. Within the same category, the videos would likely have greater commonality. This result differs from what the team initially expected for the 'Music' category to be one of the features selected since music videos on the platform tend to have the highest view count.

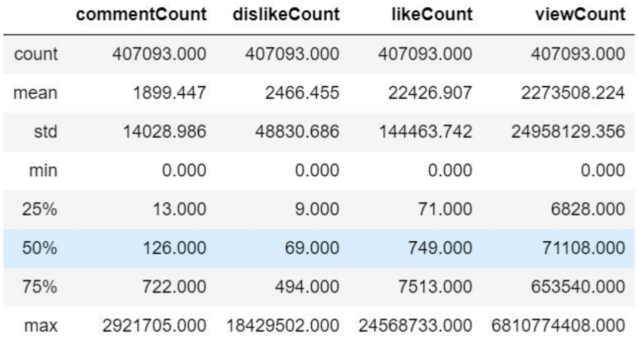
Besides that, out of the 3 sentiments a video title can have, the negative sentiment feature was selected. This result implies that the sentiment type of a video title is an important feature in determining the view count, which was in line with what the team expected.

Lastly, out of all the 17 categories of PoS tag of the title's first word, the interjection feature was selected. An interjection is a word that expresses a strong emotion. Some examples include 'wow', 'ouch', 'hurray'. This feature was likely selected because using interjections in the first few words of a video title is effective in capturing the audience's attention and increasing view count.

7. Model Building

A total of four models were built for this project—Polynomial Regression, Random Forest, Gradient Boosting and Artificial Neural Network (ANN). The dataset was split with a train-test ratio of 7:3.

For models such as Polynomial Regression and Neural Network that use distance-based methods, feature scaling is required to allow better model performance. This additional step to normalise numerical values is especially essential for values of features that vary a lot, which can be observed in Figure 22 below, where the range of the numerical data is broad, with figures spanning from 1 to 10 digits for one variable.



	commentCount	dislikeCount	likeCount	viewCount
count	407093.000	407093.000	407093.000	407093.000
mean	1899.447	2466.455	22426.907	2273508.224
std	14028.986	48830.686	144463.742	24958129.356
min	0.000	0.000	0.000	0.000
25%	13.000	9.000	71.000	6828.000
50%	126.000	69.000	749.000	71108.000
75%	722.000	494.000	7513.000	653540.000
max	2921705.000	18429502.000	24568733.000	6810774408.000

Figure 22: Summary of Numerical Variables

To tune the hyperparameters used for the model, the *GridSearchCV* Python library was used for most of the models. This was achieved by looping through a list of predefined hyperparameters, fitting the model with the training dataset and using the parameters for the best performing model.

7.1 Polynomial Regression

Polynomial Regression is a form of linear regression that can fit a nonlinear relationship between the independent and dependent variables, modelled in *n*th degree polynomial (Savaram, 2019). It is better than a regular linear regression when the relationship between the independent and dependent variables is nonlinear, which is very likely in the context of this project. The model was built using the *Pyspark MLlib* Python library, and the third-degree was applied to the model.

For polynomial regression, the *ML Tuning* Python library, which has similar functions to *GridSearchCV*, was used to perform cross-validation and grid search to tune the hyperparameters.

7.2 Random Forest

Random Forest is an ensemble machine learning algorithm of decision trees where several individual decision trees are bagged together, and the average prediction is taken (Chauhan, n.d.). Initially, each tree in the random forest will have high variances after being trained on a random sample of the training data. However, bagging the individual trees together will give the entire forest a lower variance by taking the average of the individual variances. This helps to prevent overfitting and reduces the errors of the model, compared to a single decision tree (Drakos, 2019). The model was built using *RandomForestRegressor* from the *Sklearn* Python library.

7.3 Gradient Boosting

Gradient Boosting is an ensemble machine learning algorithm of decision trees where several individual decision trees are boosted and built in a gradual, additive and sequential manner. The model relies on the intuition that the best possible next model minimises the overall prediction error when combined with the previous models (Hoare, n.d.). The model was built using the *XGBoost* Python library.

7.4 Artificial Neural Network (ANN)

Neural networks are multi-layer networks of neurons that can recognise underlying patterns and relationships in data, through a similar way like how the human brain operates. There is a broad spectrum of neural networks, ranging from Recurrent Neural Network (RNN) to Convolutional Neural Network (CNN). Out of these, Artificial Neural Network was selected for this project as it is the most suitable for tabular data (Pai, 2020).

The neural network was compiled with Mean Squared Error and the 'adam' optimiser; it ran for 100 epochs with a batch size of 1,000. The structure of the neural network is as described:

- Input layer: 1 layer with 128 neurons and 'relu' activation
- Hidden layer: 3 layers with 32 neurons each and 'relu' activation
- Output layer: 1 layer with 1 neuron and 'linear' activation

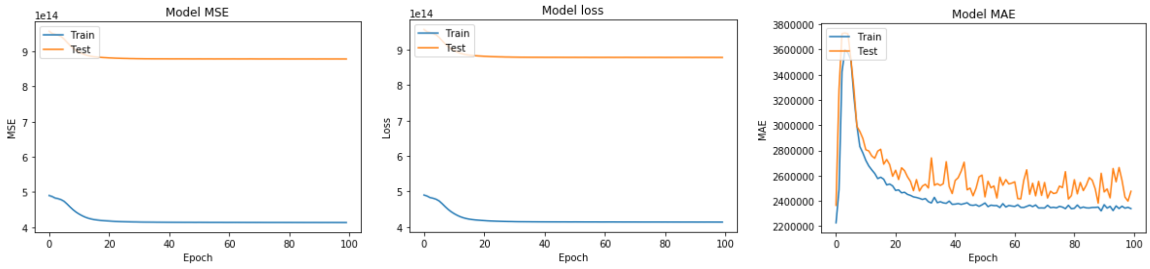


Figure 23: Artificial Neural Network Model Training History

Figure 23 above shows the training history of the ANN model. It can be observed that around the 15th epoch, the MSE and loss start to plateau, indicating that the number of epochs could have been reduced. Another interesting observation is that the validation mean absolute error (MAE) of the model fluctuates significantly. This was likely due to the fact that the model was compiled with MSE and not MAE; thus, the MAE was not taken into account during model training. After the 20th epoch, the MAE starts to plateau and fluctuates around a similar range as the number of epochs increase.

8. Model Evaluation & Selection

The models built were then evaluated based on the following metrics.

1. **Root Mean Squared Error (RMSE)** is the square root of the average squared differences between the predicted and actual observations. RMSE assigns a higher weight to larger errors; thus if large errors are present, it would affect the performance of the model. The lower the value is, the better the performance of the model.
2. **Mean Squared Error (MSE)** measures how close a fitted line is to the data points. It takes the distance between the data point and the fitted line of the model and squares them. A larger MSE signifies that the data is dispersed widely around its mean, while a smaller MSE infers otherwise. Generally, a lower MSE is preferred.
3. **Mean Absolute Error (MAE)** takes the average of the absolute differences between the predicted and actual observations. It measures the average magnitude of the errors regardless of their direction by taking the absolute values. The lower the value is, the better the performance of the model.

The summary results for the 4 models built can be seen in Table 5 below.

Model	RMSE	MSE	MAE
Polynomial Regression	24,486,700	599,598,500,178,391	2,410,707
Random Forest	29,491,373	869,741,069,564,456	2,369,201
Gradient Boosting	29,839,713	890,408,446,899,621	2,363,009
Artificial Neural Network	**20,330,611**	**413,333,730,000,000**	**2,338,276**

Table 5: Model Results

Out of all the models built, Artificial Neural Network (ANN) is the best performing model with the lowest RMSE, MSE and MAE. This may be due to the ability of neural networks in detecting complex non-linear relationships between variables.

Despite that, the MAE for ANN is still a large number, standing at a whopping 2,338,276. This can be interpreted as having an average error of 2.3 million views for the predicted view count of each video. This phenomenon can be attributed to the presence of outliers and the skewed distribution in the actual view count values, which can cause the model trained to be biased. From Figure 22 above (under Section 7), it can be observed that the standard deviation of the actual view count is over 20 million and the distribution of the statistic is skewed to the right, which supports this argument.

Other models built had similar but relatively higher metric values as compared to the ANN model, which had a significantly lower MSE of almost half the values of that for Random Forest and Gradient Boosting. A much lower MSE means that predictions by the ANN model have lower variance and generalises better compared to the other models.

9. Conclusion

As aforementioned, the hypotheses for this project are:

A. "A positive sentiment title in a YouTube video will affect views"
B. "Having humans in the thumbnail images of a YouTube video will affect views"

The models were built with the selected features based on the feature selection conducted, which included the number of humans in the thumbnail images. Thus, for the final Artificial Neural Network (ANN) model selected, it can be concluded that Hypothesis A is rejected while there is insufficient evidence to reject Hypothesis B.

Although a positive sentiment title may not have an impact on the view count of a video, the title is still an important feature of a YouTube video. Having a negative sentiment title as a selected feature can imply that it is more significant in determining view count, compared to a neutral or positive sentiment. Therefore, the sentiment of the video titles should not be discounted.

On the other hand, even though having humans in the thumbnail images can impact the view counts of a video, existing object detection models still have room for improvements in identifying the presence of humans, as mentioned in Section 4.3. In the future, the usage of a better model for feature extraction can potentially help to improve the accuracy of the model built.

Future improvements for this project can be to include a broader spectrum of YouTube data. As this project only focuses on a small subset of YouTube data, it is impossible to deduce whether a positive sentiment title or the number of humans in thumbnail images can truly affect the view count of videos in the real world for all videos on the platform. By having a more representative data, the model will then be able to achieve a more accurate prediction.

10. References

Aslam, S. (2020). YouTube by the Numbers (2020): Stats, Demographics & Fun Facts. Retrieved 16 November 2020, from https://www.omnicoreagency.com/youtube-statistics/

Bates, P. (2019). What Does This Emoji Mean? Emoji Face Meanings Explained. Retrieved 16 November 2020, from https://www.makeuseof.com/tag/emoji-english-dictionary-emoji-faces-meaning-explained/

Berman, M. (2020). Which is More Important: YouTube Views or YouTube Likes?. Retrieved 21 November 2020, from https://programminginsider.com/which-is-more-important-youtube-views-or-youtube-likes/

Chauhan, N. Decision Tree Algorithm, Explained. Retrieved 22 November 2020, from https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html

Drakos, G. (2019). Random Forest Regressor explained in depth. Retrieved 22 November 2020, from https://gdcoder.com/random-forest-regressor-explained-in-depth/

Google Developers. API Reference | YouTube Data API. Retrieved 16 November 2020, from https://developers.google.com/youtube/v3/docs

Hoare, J. Gradient Boosting Explained - The Coolest Kid on The Machine Learning Block | Displayr. Retrieved 22 November 2020, from https://www.displayr.com/gradient-boosting-the-coolest-kid-on-the-machine-learning-block/

Pai, A. (2020). ANN vs CNN vs RNN | Types of Neural Networks. Retrieved 21 November 2020, from https://www.analyticsvidhya.com/blog/2020/02/cnn-vs-rnn-vs-mlp-analyzing-3-types-of-neural-networks-in-deep-learning/

Roy, B. (2020, April 07). All about Feature Scaling. Retrieved November 17, 2020, from https://towardsdatascience.com/all-about-feature-scaling-bcc0ad75cb35

Savaram, R. (2019). Polynomial Regression. Retrieved 22 November 2020, from https://mindmajix.com/polynomial-regression

11. Appendix

Appendix A: Codes and Documentation

The codes and documentation for this project can be accessed via the following GitHub repository: https://github.com/gohjiayi/youtube_analysis.

Appendix B: Data Description

The data that was extracted from the data collection stage can be seen in Tables 6 to 8 below.

Name	Description
channelId	Unique ID of channel
channelTitle	Title of channel
category	Category of channel

Table 6: Data from Social Blade

Name	Description
channelId	Unique ID of channel
publishedAt	The date and time that the channel was created
region	The country of the channel
description	The description of the channel
viewCount	The number of times the channel has been viewed
subscriberCount	The number of subscribers that the channel has
videoCount	The number of videos uploaded to the channel

Table 7: Channel Data from YouTube API

Name	Description
channelId	Unique ID of channel
videoId	Unique ID of video
videoTitle	The video's title
description	The description of the video
thumbnails	URL to the video’s thumbnail image
publishedAt	The date and time that the video was uploaded
likeCount	The number of users who have indicated that they liked the video
dislikeCount	The number of users who have indicated that they disliked the video

commentCount	The number of comments for the video
viewCount	The number of times the video has been viewed
favouriteCount	The number of users who currently have the video marked as a favorite

Table 8: Video Data from YouTube API

Appendix C: Part-of-Speech (PoS) Tags

Table 9 below shows the coarse-grained PoS tags used.

PoS tags	**Description**	**Examples**
ADJ	adjective	big, old, green, incomprehensible, first
ADP	adposition	in, to, during
ADV	adverb	very, tomorrow, down, where, there
AUX	auxiliary	is, has (done), will (do), should (do)
CCONJ	coordinating conjunction	and, or, but
DET	determiner	a, an, the
INTJ	interjection	psst, ouch, bravo, hello
NOUN	noun	girl, cat, tree, air, beauty
NUM	numeral	1, 2017, one, seventy-seven, IV, MMXIV
PART	particle	's, not,
PRON	pronoun	I, you, he, she, myself, themselves, somebody
PROPN	proper noun	Mary, John, London, NATO, HBO
PUNCT	punctuation	., (,), ?
SCONJ	subordinating conjunction	if, while, that
SYM	symbol	$, %, §, ©, +, −, ×, ÷, =, :), 😝
VERB	verb	run, runs, running, eat, ate, eating
X	other	sfpksdpsxmsa

Table 9: Part-of-Speech (PoS) Tags

Appendix D: Object Detection Model

The YOLO v3 model is able to detect the following 80 categories of objects:

'person', 'bicycle', 'car', 'motorcycle', 'airplane', 'bus', 'train', 'truck', 'boat', 'traffic light', 'fire hydrant', 'stop sign', 'parking meter', 'bench', 'bird', 'cat', 'dog', 'horse', 'sheep', 'cow', 'elephant', 'bear', 'zebra', 'giraffe', 'backpack', 'umbrella', 'handbag', 'tie', 'suitcase', 'frisbee', 'skis', 'snowboard', 'sports ball', 'kite', 'baseball bat', 'baseball glove', 'skateboard', 'surfboard', 'tennis racket', 'bottle', 'wine glass', 'cup', 'fork', 'knife', 'spoon', 'bowl', 'banana', 'apple', 'sandwich', 'orange', 'broccoli', 'carrot', 'hot dog', 'pizza', 'donut', 'cake', 'chair', 'couch', 'potted plant', 'bed', 'dining table', 'toilet', 'tv', 'laptop', 'mouse', 'remote', 'keyboard', 'cell phone', 'microwave', 'oven', 'toaster', 'sink', 'refrigerator', 'book', 'clock', 'vase', 'scissors', 'teddy bear', 'hair drier', 'toothbrush'