Analysis of Likes and Dislikes on YouTube Videos

Divya Khairnar (dk6350@rit.edu)

Table of Contents

Section 1 – Introduction

YouTube has become one of the most popular social media used all over the world. It has over 2.7 billion active users making it the second-biggest social media platform [1]. This platform provides social media influencers and content creators an easy way to reach out to the audience. As the second most used social media platform in the world, predicting likes and dislikes poses a significant challenge. Moreover, the geographically spread users add complexity to the analysis [2]. The increasing number of videos uploaded can be observed in Figure 1, more than 400 hours of videos are uploaded every minute [3] [4]. Nowadays, creating content on YouTube has become a full-time job opportunity, users can earn money when they reach a certain number of viewers and likes. Hence, analyzing and predicting these factors is important for understanding the dynamics of content performance and engagement.

Understanding the audience's likes and dislikes is very crucial for content managers, social media influencers, and content creators. This is because understanding these metrics can directly impact their video visibility and eventually help in revenue generation. As YouTube has a large volume of users and has a global reach with content uploaded almost every minute, the ability to predict likes and dislikes can help creators in data-driven decision making for content optimization. Implementing a machine learning prediction model will help to identify key attributes affecting the reactions on videos. Furthermore, creating visualizations to provide insights into the YouTube engagement can make it easier to communicate outcomes with creators having non-technical background.

This report includes various sections covering prior work, methodology and results. The prior work section includes research of various papers to study the methods of analysis and prediction of data. The methodology section outlines the approach used to data cleaning, analysis, visualization, and prediction of engagement factors – likes and dislikes, followed by experiments and results section to demonstrate the findings. Lastly, the report concludes with a discuss of key findings and future work.

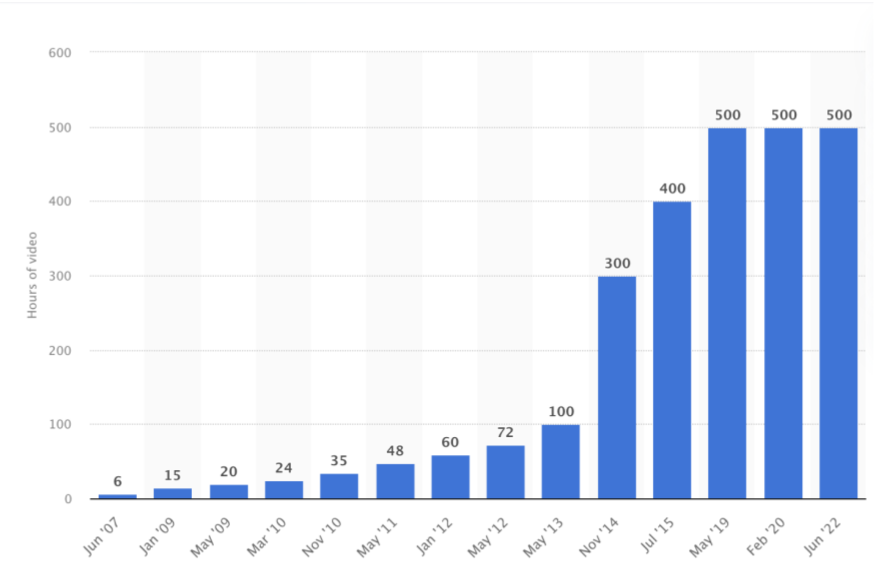


Figure 1. Hours of video uploaded per minute [3] [4]

Section 2 – Prior Work

The attributes such as the category of video and geographic location were studied by J Fang et.al to study the preference for YouTube videos [2]. They utilized linear regression and neural network to examine the video preference and used the R-square metrics to compare the models [2]. The paper focuses on only three attributes to analyze the video whereas attributes such as likes and dislikes can also be explored to study the preference of the video and studying more attributes can help the model to predict better. Additionally, only two models are implemented which makes it difficult to compare the performance of the model. Model evaluation is easier when there are more models as every model performs differently with the same dataset and relying on a single metric for comparing performance can give us an incomplete understanding of the models.

Big data analysis was conducted by Johanes Fernandes et.al. on YouTube data [5]. The authors utilized Tableau to analyze YouTube data for five countries which are – the USA, France, Russia, India, and Japan [5]. In this analysis, the authors have utilized one year's worth of data i.e., from 2017 to 2018, which I believe is insufficient for concluding a comprehensive analysis.

In 2023, M. Omkar et.al performed predictive modeling of diabetes hospital readmission using machine learning algorithms [6]. The predictive analysis is mainly conducted to improve healthcare conditions by predicting and analyzing the diabetes in an individual [6]. The authors have used AWS EC2 in their application but can cost a lot when the server is run continuously, hence additional monitoring of the servers is required [6]. The authors have used only accuracy as the performance metrics, and the accuracy of the models is very high which indicates that there might be overfitting issue and presence of imbalance in the dataset.

A model was developed by H. Bhatta et.al to predict the popularity of YouTube videos using the viewer engagement feature [7]. The engagement of the video was calculated by a popularity parameter which was obtained by dividing the total number of likes by the total number of views [7]. The biggest drawback of this paper is that a high engagement ratio will incline towards the higher popularity of the video. Hence using the engagement ratio might not be the best attribute to study the popularity of the video. Furthermore, the authors have used only a single category of video for their study, using more categories will allow a more generalized prediction model and provide an overall view.

M. P et.al have presented a research paper on “Prediction of YouTube View Count using Supervised and Ensemble Machine Learning Techniques” in the year 2022 [8]. The authors have used various supervised and ensemble machine learning techniques, including Multiple Linear Regression (MLR), Random Forest Regressor (RFR), Decision Tree Regressor (DTR), XGBoost Regressor (XGB), and Gradient Boost Regressor (GBR) to predict the view count of YouTube videos and have used various metrics such as Mean Absolute Error(MAE), Mean Squared Error(MSE), Root Mean Squared Error(RMSE), and R-squared(R2) [8]. Although they have used multiple models and metrics for comparison which gives a better understanding of how models are performing, the values of MSE, RMSE, and MAE are quite high. This indicated that there are large errors in the predictions made by the models, and the model has a poor fit. Hence model tuning and feature engineering will be necessary to ensure accurate prediction of view count.

Prediction of ratings of trending YouTube videos was conducted by G. Gupta et.al using machine learning algorithms [9]. The authors have utilized multiple machine learning models such as

Random Forest, Decision Tree, Support Vector Machine, Logistic Regression, and K-Nearest Neighbor to predict the ratings and have used accuracy to compare the performance of the metrics. Not only they have used only one metric for model comparison, but the value of accuracy ranges from 97 – 100% which is quite high and indicates a major overfitting issue. The dataset used in this research is limited and has only 999 fields, which may fail to capture a wide range of content. Expanding the dataset, implementing measures to handle overfitting issues, and using more metrics to compare the model can be beneficial in this case.

After reviewing the above findings, it can be said that there is still some scope for improvement in the field of YouTube Data Analysis, to predict the likes and dislikes of the videos. One of the crucial points is to deal with overfitting issues in the dataset. The proposed system will also use multiple machine learning models for better comparison and will also deal with class imbalance. The use of evaluation matrices Mean Absolute Error(MAE), Mean Squared Error(MSE), Root Mean Squared Error(RMSE), and R-squared(R2) will help to better compare the outcomes of all the prediction models, and a best-fit model can be determined. Finally, using Tableau for data visualization can help better understand the outcomes and factors affecting likes and dislikes for a video.

Section 3 – Methodology

This section outlines the steps used to predict the likes and dislikes of YouTube videos for datasets from India and USA. As shown in Figure 2, the methodology consists of six stages - data exploration, data pre-processing, feature engineering, feature selection, model implementation, and dashboard development.

STAGE 1 – DATA EXPLORATION
STAGE 2 – DATA PREPROCESSING
STAGE 3 – FEATURE ENGINEERING
STAGE 4 - FEATURE SELECTION
STAGE 5 – MODEL IMPLEMENTATION
STAGE 6 – DASHBOARD DEVELOPMENT

Figure 2. Stages of Methodology

3.1 Data Exploration

The YouTube video dataset used in this project is sourced from Kaggle [10]. The dataset consists of trending video data from 11 countries from which we are going to use two countries India and USA. The attributes in the dataset include video title, channel title, publish time, tags, views, likes, dislikes, description, and comment count. In total the dataset includes 268787 entries for the US dataset and 251277 entries for India dataset.
Various visualizations were created to gain insights into relation of various attributes with the engagement attributes as seen below:

- Figure 3. shows the relation between videos popularity measured by the view count and likes and dislikes. It can be observed that the video with higher view count correspond to more likes, hence it can be said that popular videos have received positive feedback from the audience. But the figure also shows that dislikes on a video also increases with popularity.

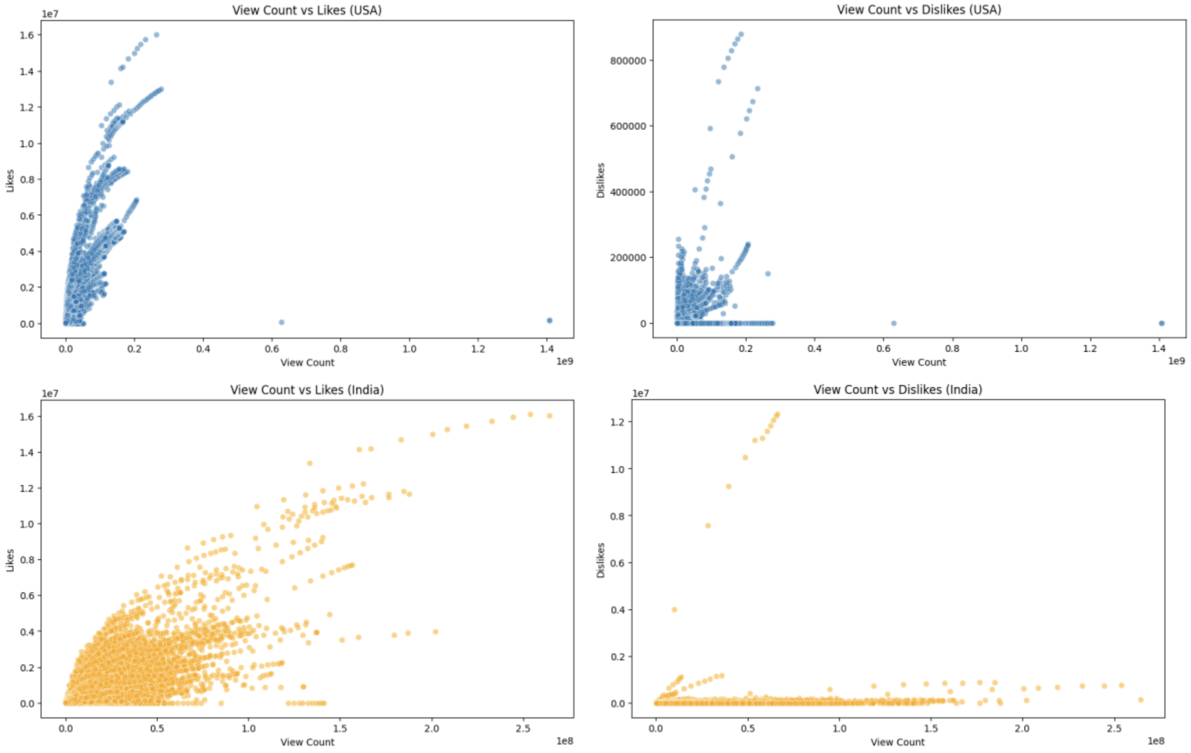


Figure 3. View Count vs Likes and Dislikes

- The bar graph in Figure 4 Shows the distribution of videos in each category for both USA and India. In both the countries ‘Entertainment’ has the highest number of records. USA has the lowest record for ‘People and Blogs’ category whereas India has the lowest records for ‘Gaming’ category. It can be said that the categories with lowest records may have less competition or low growth rate.

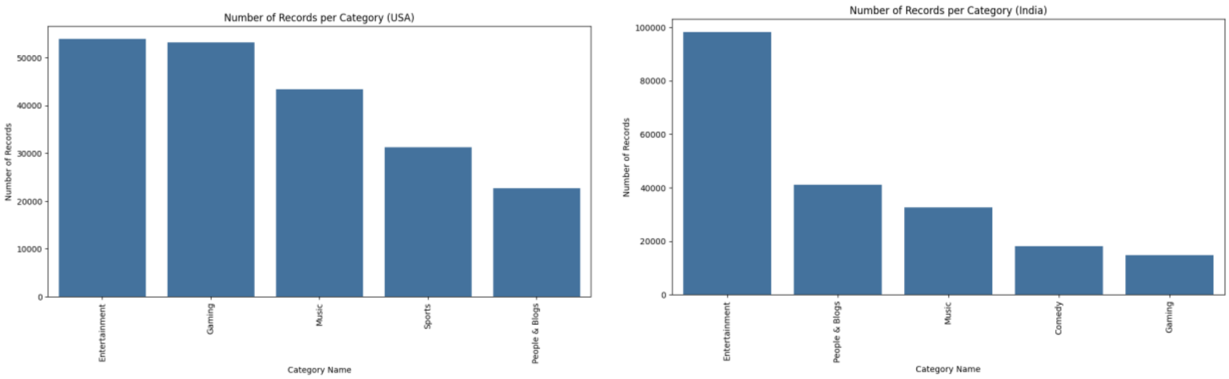


Figure 4. Records per category for India and USA

- The average view count, likes, and dislikes as per category is illustrated in the bar plot shown in Figure 5 and Figure 6 The visualization offers analysis of how engagement is affected by different categories. In general, for both the countries ‘Music’ has the highest number of view count, likes and dislikes.

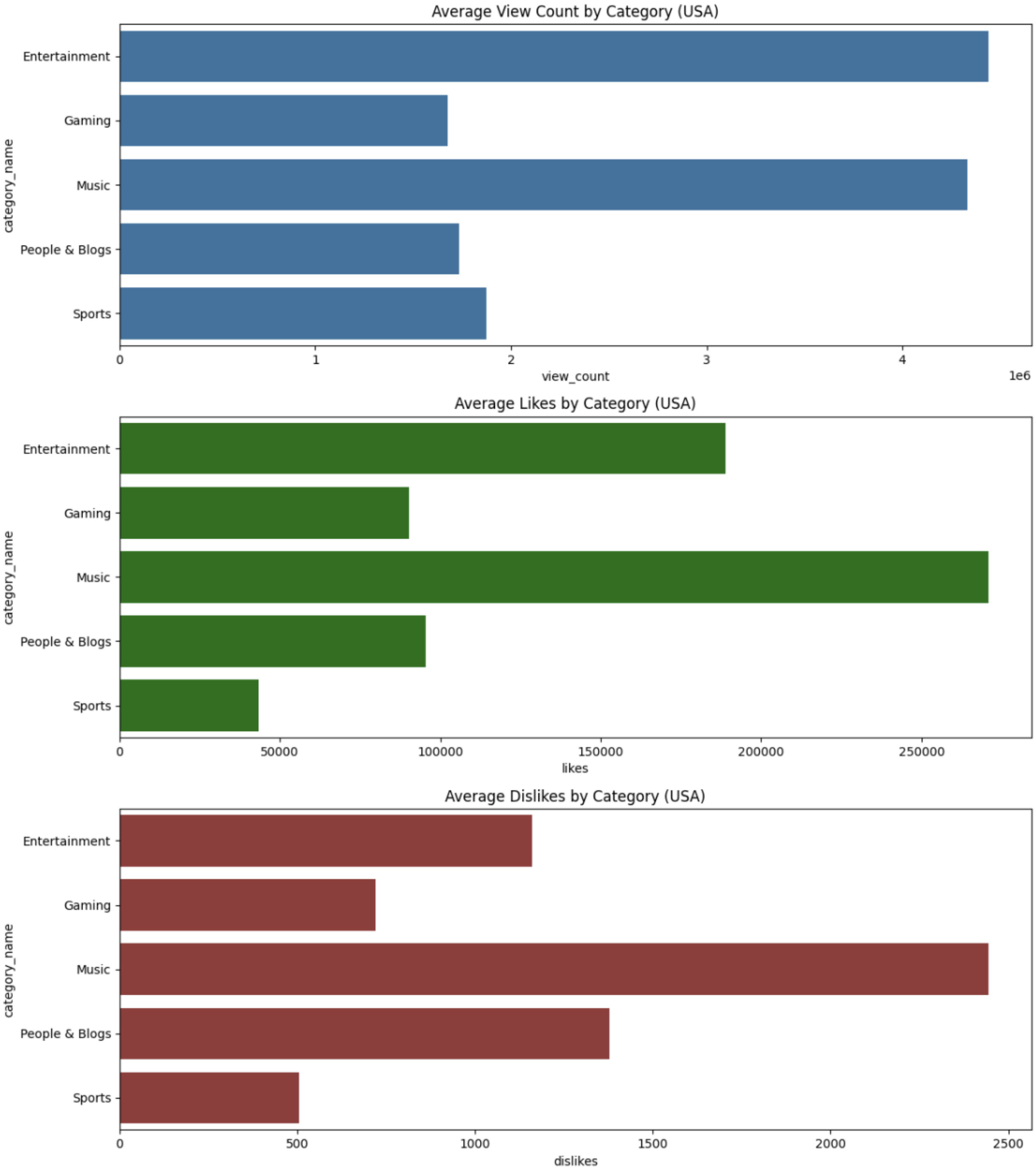


Figure 5. Average view count, likes and dislikes per category(USA)

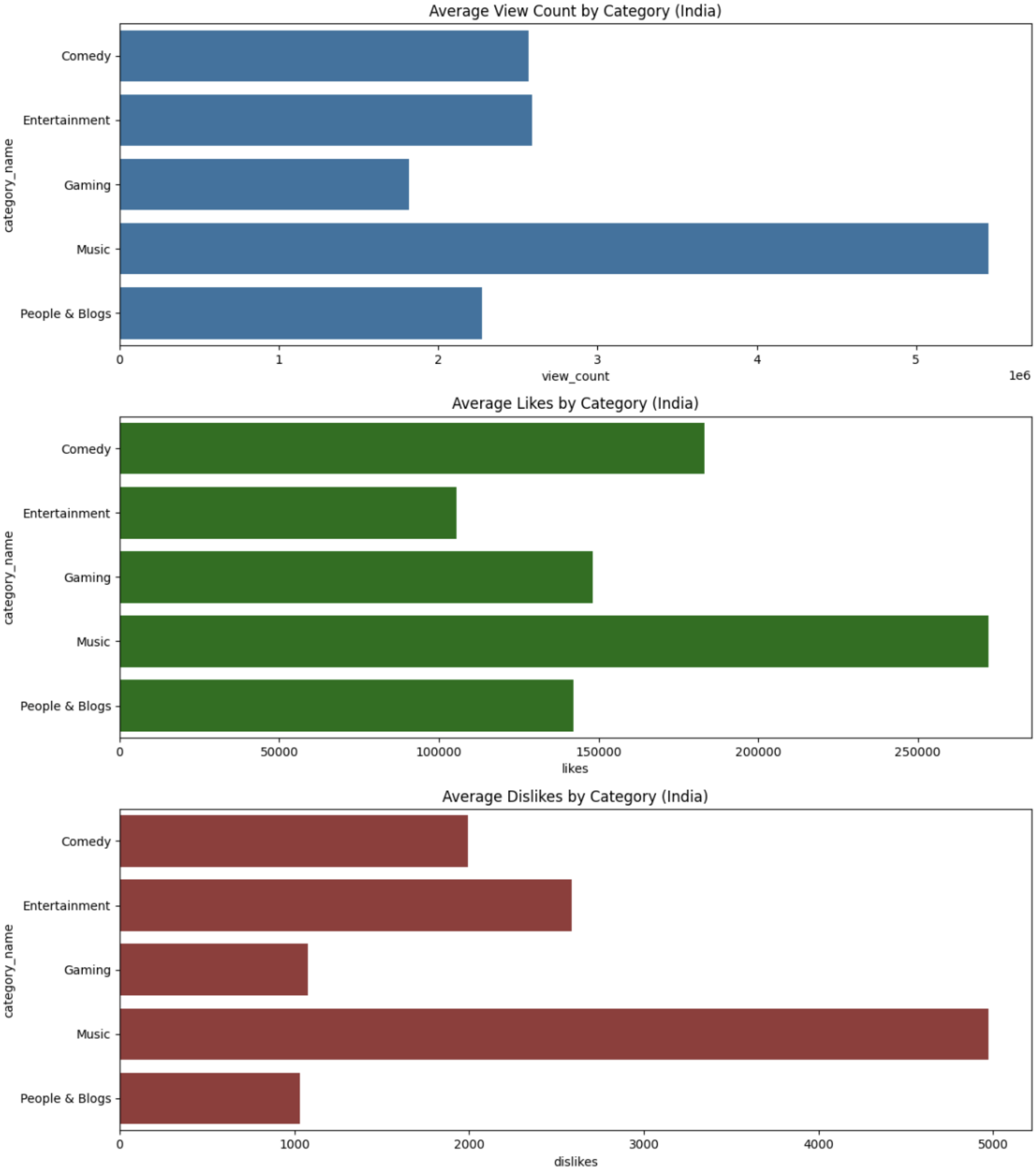


Figure 6. Average view count, likes and dislikes per category(India)

- The line plot in Figure 7 represents the user engagement by hour of the day. The peak in the line plot is when the video receives the greatest number of likes or dislikes. For US, it can be observed that the peak is higher during the early hours of the day. Whereas for India, the peak can be notices during late night hours.

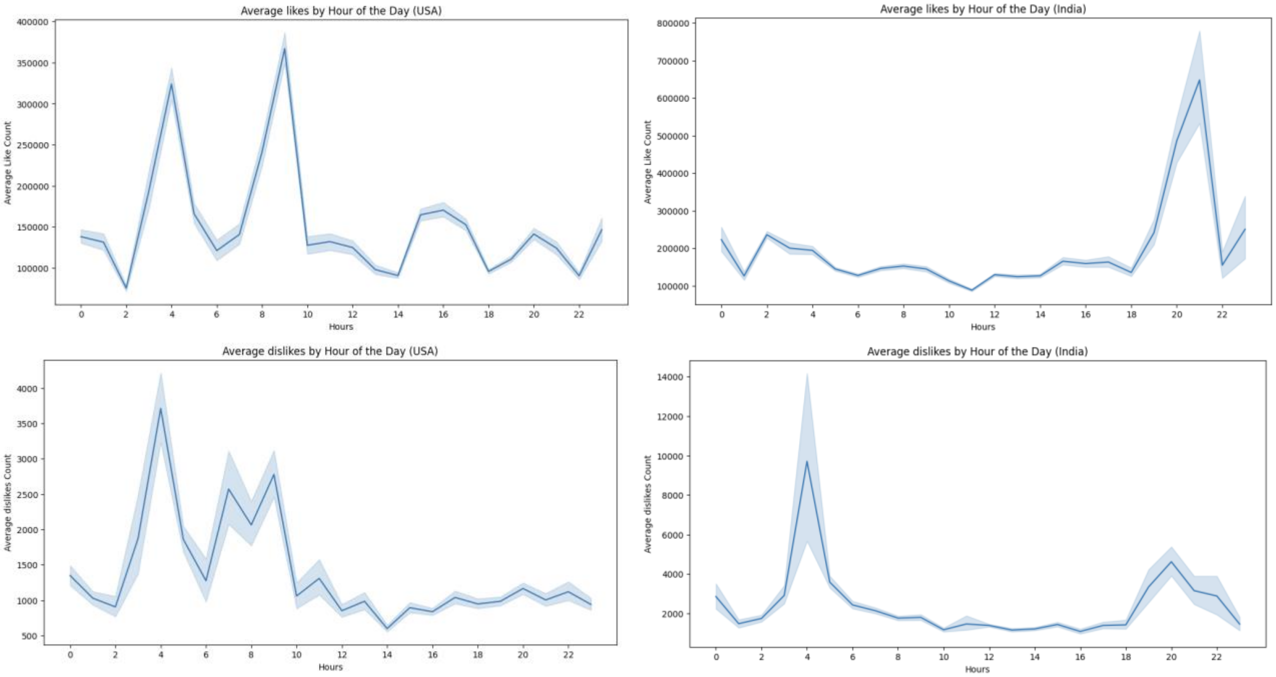


Figure 7. Average likes and dislikes by hour of the day

- The most popular tags are represented via word cloud as shown in Figure 8. The word cloud highlights the most frequent words from the tags attribute showing the popular topics used for videos in both countries

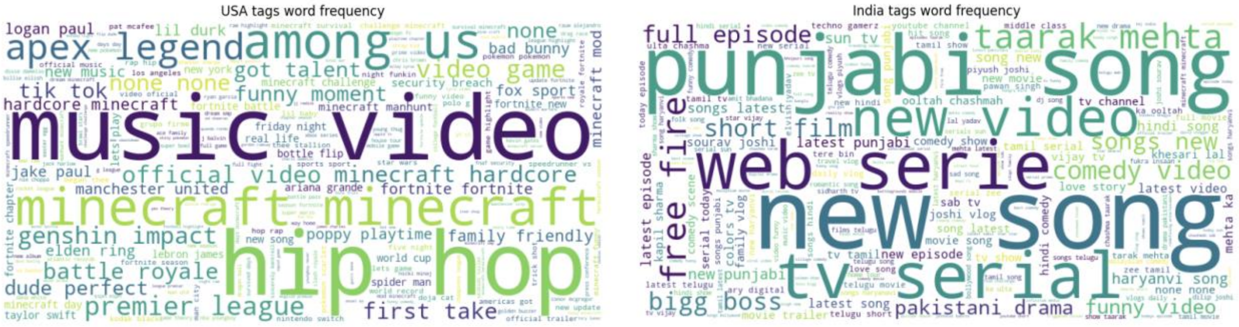


Figure 8. Word cloud

- Figure 9 illustrates audience engagement according to the day of the week. It can be seen that the average likes are consistent during the weekdays and increases during the weekend. For dislikes in USA the number is average during entire week, but a spike is noticed during Friday. The same spike in average number of dislikes is noticed during Wednesday in India.

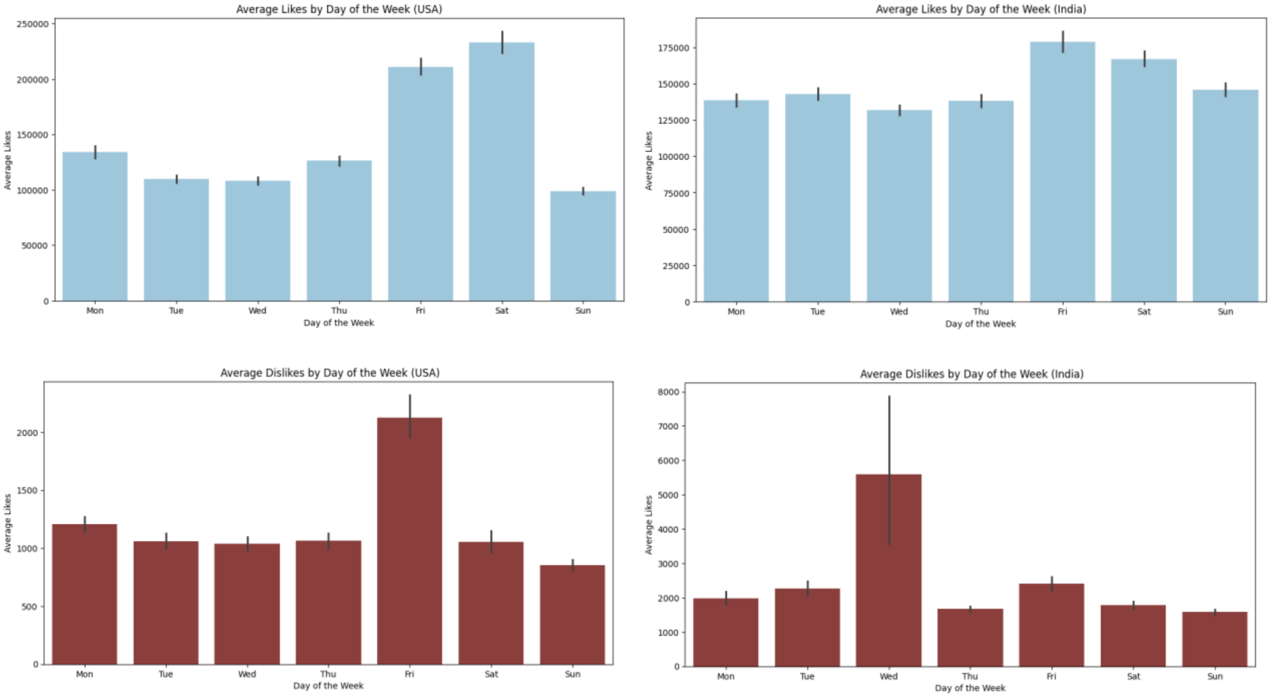


Figure 9. Average likes and dislikes by day of the week

3.2 Data preprocessing

In this step it is made sure that the data is ready for further analysis. Since the datasets had no null values hence there was no need to handle missing values. Duplicate values from the datasets were deleted to maintain data integrity. Top 5 categories with the highest number of records were focused for further analysis out of the 15 categories. Text cleaning was performed on the tags column to remove special characters and the ‘|’ separators between the tags were removed. All the tags were converted to lowercase for uniformity. Lastly, StandardScaler was used from scikit-learn library to standardize the attributes of the dataset by scaling them to have a mean of 0 and a standard deviation of 1. This ensures all the attributes have a uniform scale and helps to contribute to the prediction models equally.

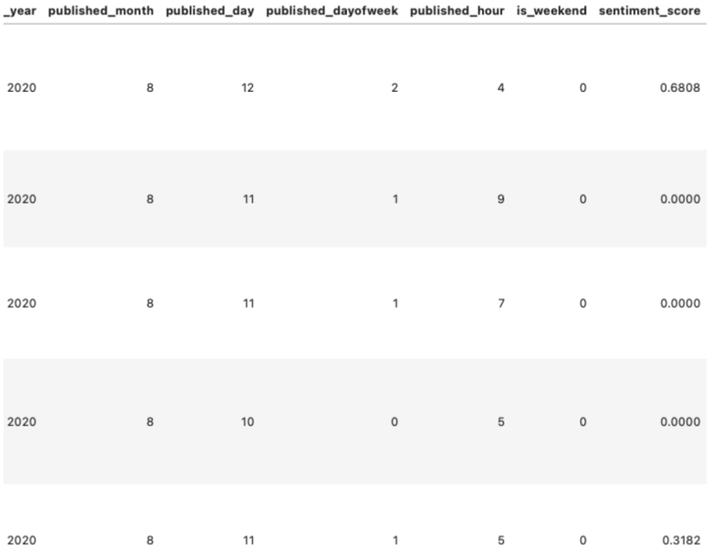
3.3 Feature engineering

In the third stage, several new attributes were derived from the existing attribute to improve the accuracy of prediction and provide deeper insights into likes and dislikes of the videos as shown in Figure 10. The derived attributes are as follows:

- From the “publishedAt” attribute various temporal date-time attributes were extracted such as-
  a. Published year (Ranges from 2020 - 2024)
  b. Published month - ranging from 1 (January) to 12 (December)

c. Published day - ranging from 1 to 30
d. Published day of the week - ranging from 0 (Monday) to 6 (Sunday)
e. Published hour – ranging from 0 to 23
f. Is weekend – a binary attribute where weekday is represented by 0 and weekend by 1

- Sentiment score – Sentiment analysis was performed using the python nltk library on the tags column and sentiment score was generated for each video. The sentiment score ranges from -1 (most negative) to +1(most positive), the score 0 indicates neutral tags. The sentiment score helps to understand how the mood of the video affects viewer engagement.



_year	published_month	published_day	published_dayofweek	published_hour	is_weekend	sentiment_score
2020	8	12	2	4	0	0.6808
2020	8	11	1	9	0	0.0000
2020	8	11	1	7	0	0.0000
2020	8	10	0	5	0	0.0000
2020	8	11	1	5	0	0.3182

Figure 10. Derived attributes

3.4 Feature selection

Certain features were selected to train the model based on how effectively they predict likes and dislikes.

- The features used to predict likes of the video are –
'view_count', 'dislikes', 'comment_count', 'categoryId', 'published_year', 'published_month', 'published_day', 'published_hour', 'published_dayofweek', 'is_weekend', 'sentiment_score'

- The features used to predict dislikes of the video are –
'view_count', 'likes', 'comment_count', 'categoryId', 'published_year', 'published_month', 'published_day', 'published_hour', 'published_dayofweek', 'is_weekend', 'sentiment_score'

3.5 Model Implementation

In the model implementation phase 4 supervised machine learning models are executed. The models are Linear Regression, Random Forest, Decision Tree, K-nearest neighbors. Based on the characteristics each model was selected for its distinct strengths in prediction of likes and dislikes of YouTube video. Machine learning library Scikit-Learn was used to implement all models as it performs effectively with real world datasets. Scikit-Learn is a user-friendly library and has a wide variety of algorithms for both supervised and unsupervised models. It also provides various preprocessing functionalities such as normalizing, scaling, and encoding.

- Linear regression – It is a kind of machine learning model which is relatively straightforward and easy to understand as it assumes a linear relationship between variables. The model computes relation between a dependent variable such as likes or dislikes and one or more independent variable such as view count, sentiment score, day, time. Despite its simplicity, it may struggle to identify patterns due to outliers or non-linearity of the data [11].

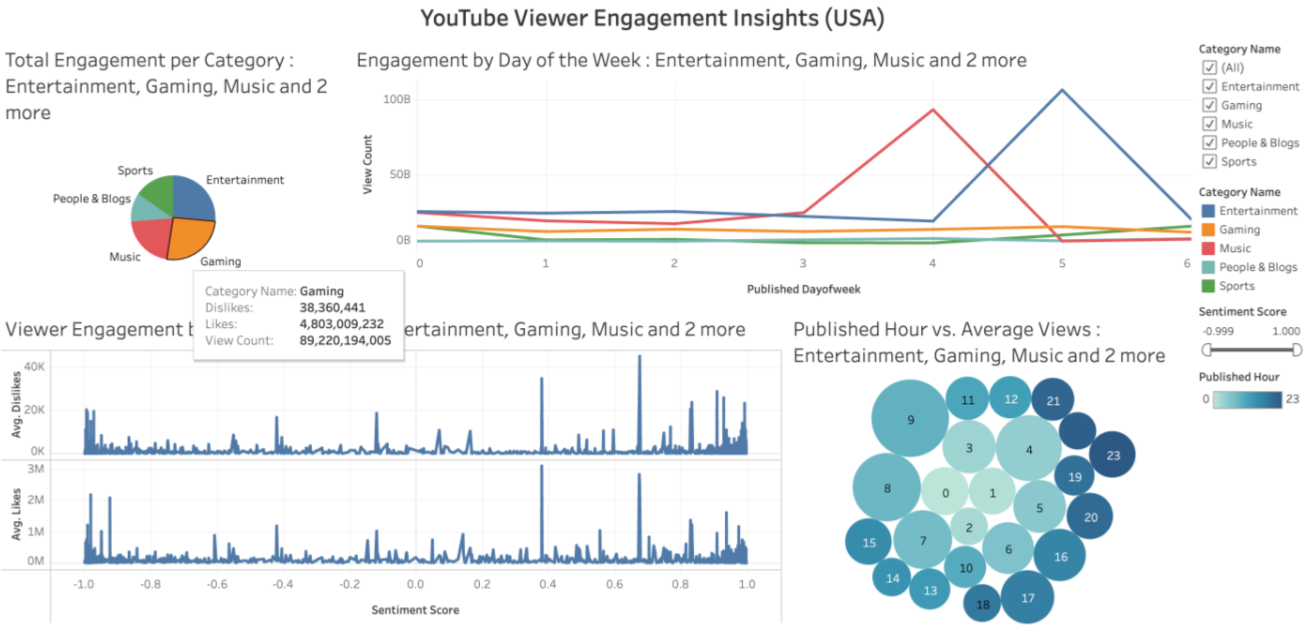
- Decision Tree - Decision Trees generate a sequence of binary splits based on feature values in order to forecast the target variable when it comes to predicting YouTube video engagement metrics (likes and dislikes). Since the decision-making process may be represented as a tree structure, one of the main benefits of decision trees is their interpretability [11]. Decision often tends to overfit when there are increase in nodes due to complexity of model but can be minimized using Random Forest model.

- Random Forest – This machine learning model performs training by creating multiple decision trees using a random subset of dataset. This randomness helps the model to increase accuracy, decrease overfitting issue and helps to increase performance with high complexity datasets [11]. Hence making the model suitable for a situation where the model needs to study complex relations between the features and target variable.

- K-Nearest Neighbor – This model predicts the target variable by using the average of the k nearest data points. KNN determines the Euclidean distance between the data points with every other point in the dataset to predict the engagement variables [11]. Standardizing the input helped to scale the features improving the performance of the model.

3.6 Dashboard Development

To present a visual representation of how various factors affect the engagement metrics of YouTube videos such as view, likes and dislikes, the Tableau dashboard was created for both US and India's dataset. As seen in Figure 10 and Figure 11, to make the exploration easy, the dashboard includes various interactive charts. The pie chart displays the overall distribution of categories such as Comedy, Entertainment, Gaming, Music, and People & Blogs. The fluctuation of count of engagement metrics by the day of the week is illustrated by a line graph. This graph shows how the count of engagement increases and decreases through the week from Monday to Sunday, helping us understand which day is the best to upload a new video. Another line chart shows the relationship between the sentiment score and the viewer engagement. This illustrates how the likes and dislikes change with the change in emotion surrounding the videos. Finally, a packed bubble chart displays the engagement according to the published hour, giving information about when viewers are most likely to participate.



Category Name: Gaming
Dislikes: 38,360,441
Likes: 4,803,009,232
View Count: 89,220,194,005

Figure 11. Tableau dashboard for US dataset

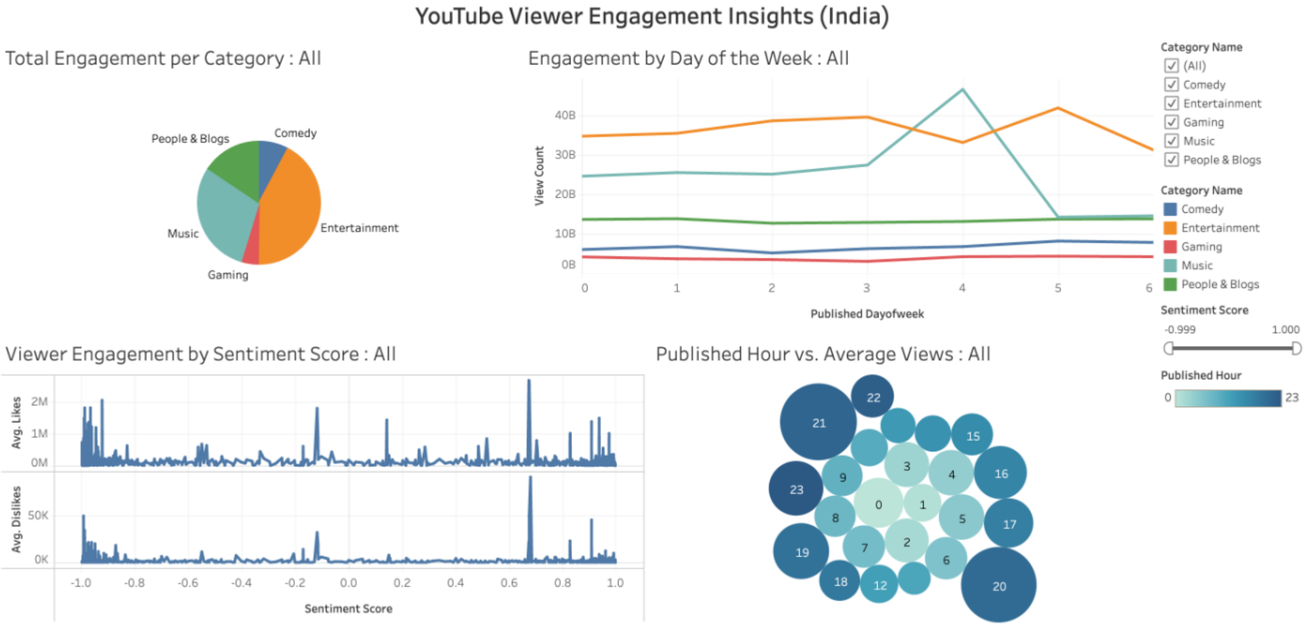
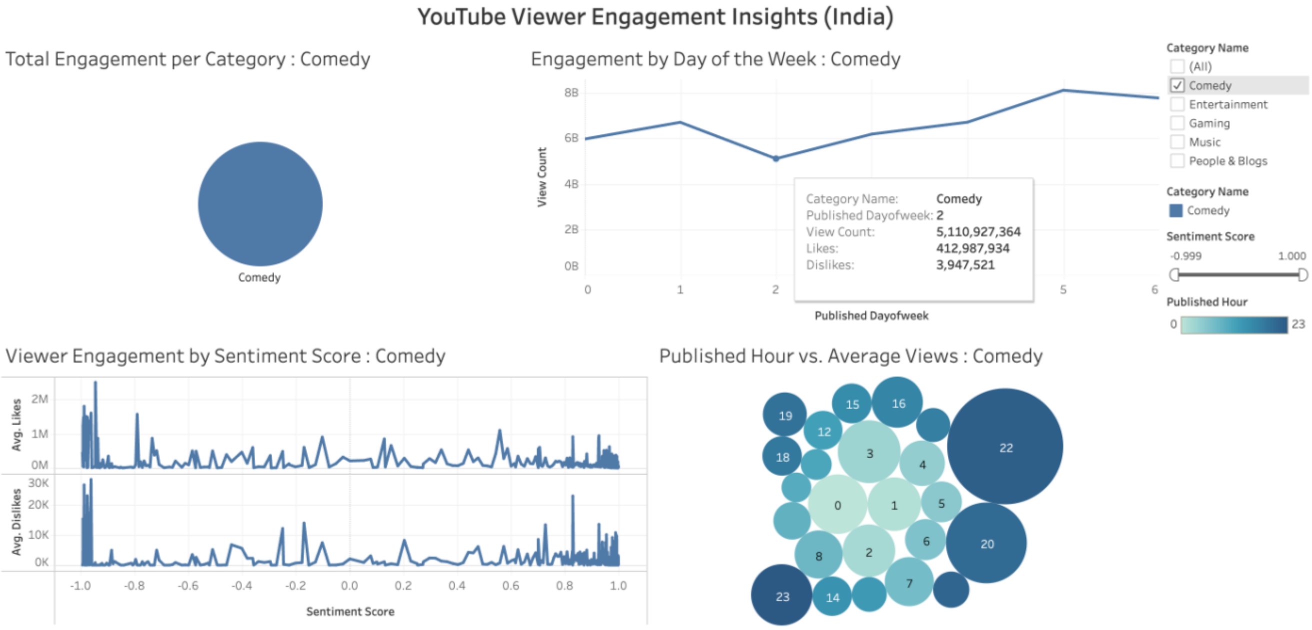


Figure 12. Tableau dashboard for Indian dataset

Additionally, the dashboard also features filters for category name and sentiment score that allows user to refine and analyze data from different perceptive. In Figure 12 It can be observed that the filter for the category 'Comedy' has been applied, hence the dashboard shows the engagement insights for the Comedy genre. This will help users who are focusing on a particular category to focus their analysis. For users focusing on more than one category can filter using the Category Name checkbox.



Category Name: Comedy
Published Dayofweek: 2
View Count: 5,110,927,364
Likes: 412,987,934
Dislikes: 3,947,521

Category Name
□ (All)
☑ Comedy
□ Entertainment
□ Gaming
□ Music
□ People & Blogs

Sentiment Score: -0.999 – 1.000

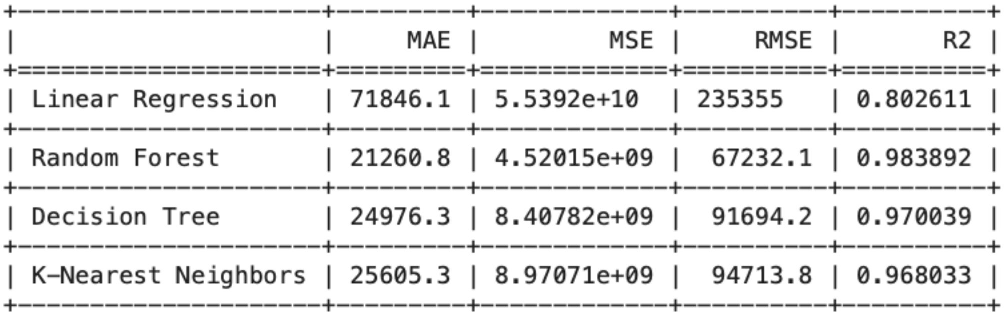
Figure 13. Tableau dashboard for comedy category (Indian dataset)

Section 4 – Experiments and Results

With a focus on likes and dislikes attributes from the US and India dataset, this section outlines the experiments conducted to the predict the engagement features for YouTube videos. This experiment involved training four machine learning models, such as Linear Regression, Decision Tree, Random Forest, and K-Nearest Neighbors, were trained on preprocessed data to analyze their performance. Metrices like Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2) were used to evaluate outcomes of each model. The aim of this experiment was to determine which model performs the best using selected features and accurately predicted the engagement features.

4.1 US likes prediction

All models were trained, and their performance were compared in order to predict the likes in US YouTube video dataset. As seen in Figure 14, with the lowest RMSE and highest R^2 values, Random Forest model performed the best. The Decision Tree and KNN models displayed a slight greater RMSE but almost similar R^2 values. On the other hand, the Linear regression model displayed comparatively greater RMSE values. This indicates that the Random Forest model could capture complex relationships in data whereas other models were not as effective for underlying patterns.

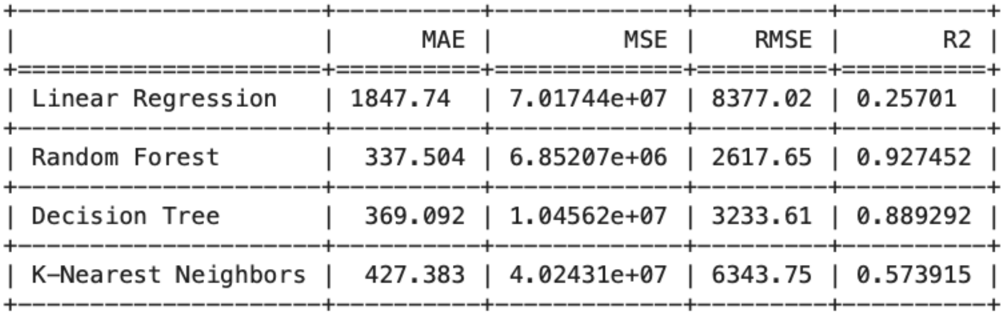


```
+---------------------+---------+-------------+----------+----------+
|                     |     MAE |         MSE |     RMSE |       R2 |
+=====================+=========+=============+==========+==========+
| Linear Regression   | 71846.1 | 5.5392e+10  | 235355   | 0.802611 |
+---------------------+---------+-------------+----------+----------+
| Random Forest       | 21260.8 | 4.52015e+09 |  67232.1 | 0.983892 |
+---------------------+---------+-------------+----------+----------+
| Decision Tree       | 24976.3 | 8.40782e+09 |  91694.2 | 0.970039 |
+---------------------+---------+-------------+----------+----------+
| K-Nearest Neighbors | 25605.3 | 8.97071e+09 |  94713.8 | 0.968033 |
+---------------------+---------+-------------+----------+----------+
```

Figure 14. US likes prediction results

4.2 US dislikes prediction

As observed in Figure 15, similar patterns were found in predicting dislikes for US dataset. Demonstrating its ability to handle non-linear relationships, Random Forest performed better than other models with 2617.65 RMSE value and 0.927452 R² value. The performance of Decision Tree model was also good with a slightly lower accuracy than that of Random Forest. In this case, Linear Regression and KNN models performed less accurately.

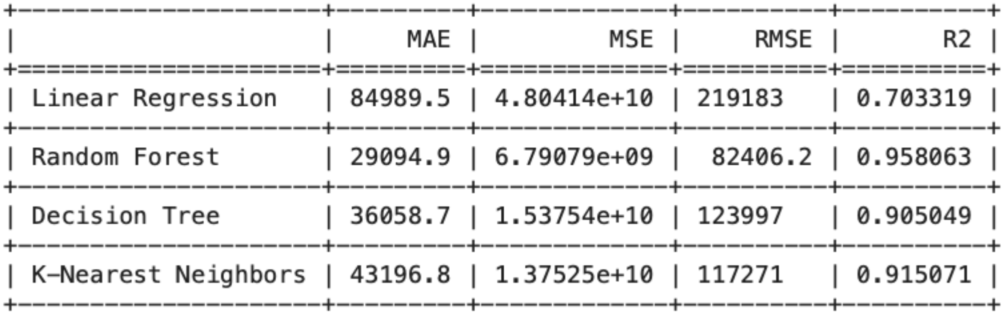


```
+---------------------+----------+-------------+---------+----------+
|                     |      MAE |         MSE |    RMSE |       R2 |
+=====================+==========+=============+=========+==========+
| Linear Regression   | 1847.74  | 7.01744e+07 | 8377.02 | 0.25701  |
+---------------------+----------+-------------+---------+----------+
| Random Forest       |  337.504 | 6.85207e+06 | 2617.65 | 0.927452 |
+---------------------+----------+-------------+---------+----------+
| Decision Tree       |  369.092 | 1.04562e+07 | 3233.61 | 0.889292 |
+---------------------+----------+-------------+---------+----------+
| K-Nearest Neighbors |  427.383 | 4.02431e+07 | 6343.75 | 0.573915 |
+---------------------+----------+-------------+---------+----------+
```

Figure 15. US dislikes prediction results

4.3 India likes prediction

Figure 16 shows that with the lowest RMSE of 82406.2 and the highest R² value of 0.958063, Random Forest was found to be the most accurate model for likes prediction in India YouTube video dataset. All the other three models which are Decision Tree, KNN and Linear Regression found it difficult to generate competitive results.

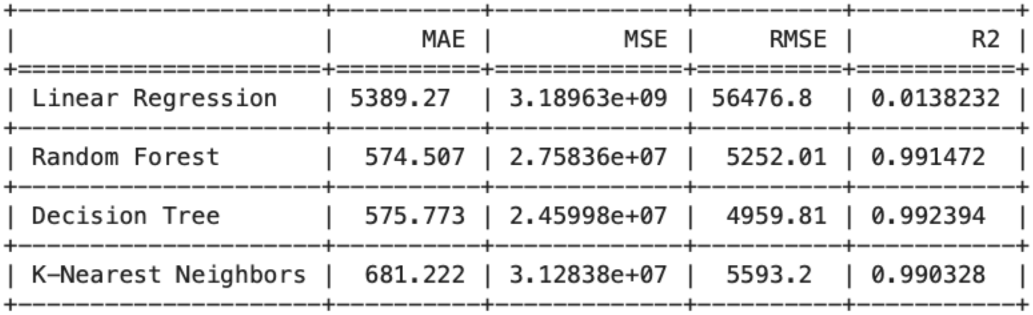


```
+---------------------+---------+-------------+----------+----------+
|                     |     MAE |         MSE |     RMSE |       R2 |
+=====================+=========+=============+==========+==========+
| Linear Regression   | 84989.5 | 4.80414e+10 | 219183   | 0.703319 |
+---------------------+---------+-------------+----------+----------+
| Random Forest       | 29094.9 | 6.79079e+09 |  82406.2 | 0.958063 |
+---------------------+---------+-------------+----------+----------+
| Decision Tree       | 36058.7 | 1.53754e+10 | 123997   | 0.905049 |
+---------------------+---------+-------------+----------+----------+
| K-Nearest Neighbors | 43196.8 | 1.37525e+10 | 117271   | 0.915071 |
+---------------------+---------+-------------+----------+----------+
```

Figure 16. India likes prediction results

4.4 India dislikes prediction

In the fourth case, which is predicting dislikes in India’s dataset the Decision Tree model performed the best and has most accurate findings with lowest RMSE and highest R². As seen in Figure 17, Random Forest model performed closely behind, with a slightly greater RMSE and almost similar R², followed by KNN which also performed well. The Linear Regression model performed badly comparing to the top tree-based algorithms indicating the inability to capture complexities in the dataset.



```
+---------------------+----------+-------------+----------+-----------+
|                     |      MAE |         MSE |     RMSE |        R2 |
+=====================+==========+=============+==========+===========+
| Linear Regression   | 5389.27  | 3.18963e+09 | 56476.8  | 0.0138232 |
+---------------------+----------+-------------+----------+-----------+
| Random Forest       |  574.507 | 2.75836e+07 |  5252.01 | 0.991472  |
+---------------------+----------+-------------+----------+-----------+
| Decision Tree       |  575.773 | 2.45998e+07 |  4959.81 | 0.992394  |
+---------------------+----------+-------------+----------+-----------+
| K-Nearest Neighbors |  681.222 | 3.12838e+07 |  5593.2  | 0.990328  |
+---------------------+----------+-------------+----------+-----------+
```

Figure 17. India dislikes prediction results

Section 5 – Conclusions and Future Work

To conclude, the objective of this project was to use a variety of machine learning models to predict the engagement metrics and to study the factors that influence the engagement of YouTube videos. One of the main findings is that ensemble approaches specifically, Decision Tree and Random Forest were more successful in forecasting engagement measures because they can identify non-linear relationships and interactions in the data. Despite being straightforward and easy to understand, linear regression proved less successful in representing the dataset's complexity. Important insights regarding viewer behavior were also uncovered by visualizing data, including how sentiment, publication timing, and video genres affect engagement. To enable them to optimize their content strategy, users were able to concentrate on specific categories and trends due to the dynamic visualization and exploration capabilities of the interactive Tableau dashboard.

Several approaches can improve the study and forecast in the future. A more comprehensive view of trends can be studied by expanding the dataset to include more countries and categories. The performance of the model can be further enhanced by adding features such as video length and video quality. Advance deep learning methods such as neural networks can be explored for more intricate pattern identification. Lastly, adding real-time data streaming to the dashboard can provide real-time insights which will help make the analysis more useful for content creators and business owners.

References

[1] Shewale, R. (2023, November 28). *YouTube statistics for 2023 (demographics & usage)*. demandsage. https://www.demandsage.com/youtube-

[2] J. Fan and T. Lian, "Youtube Data Analysis using Linear Regression and Neural Network," *2022 International Conference on Big Data, Information and Computer Network (BDICN)*, Sanya, China, 2022, pp. 248-251, doi: 10.1109/BDICN55575.2022.00055.

[3] F. Shaikh, D. Pawaskar, A. Siddiqui and U. Khan, "YouTube Data Analysis using MapReduce on Hadoop," *2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT),* Bangalore, India, 2018, pp. 2037-2041, doi: 10.1109/RTEICT42901.2018.9012635.

[4] Published by Laura Ceci, & 5, S. (2023, September 5). *YouTube: Hours of video uploaded every minute 2022*. Statista. https://www.statista.com/statistics/259477/hours-of-video-uploaded-to-youtube-every-minute/

[5] Andry, Johanes & Tannady, Hendy & Limawal, Isabelle & Rembulan, Glisina & Marta, Rustono. (2021). “BIG DATA ANALYSIS ON YOUTUBE WITH TABLEAU”. *Journal of Theoretical and Applied Information Technology*. 99. 5460-5469.

[6] M. Omkar and K. Nimala, "Machine Learning based Diabetes Prediction using with AWS cloud," *2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES),* Chennai, India, 2022, pp. 1-7, doi: 10.1109/ICSES55317.2022.9914160.

[7] H. Batta, A. V. Murthy and S. Savitri, "Predicting Popularity of YouTube videos using Viewer Engagement Features," *2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, Noida, India, 2022, pp. 77-81, doi: 10.1109/Confluence52989.2022.9734220.

[8] M. P, M. A, S. R. J, and S. N. S. K, "Prediction of YouTube View Count using Supervised and Ensemble Machine Learning Techniques," 2022 International Conference on Automation, Computing and Renewable Systems (ICACRS), Pudukkottai, India, 2022, pp. 1038-1042, doi: 10.1109/ICACRS55517.2022.10029277.

[9] G. Gupta, S. Tiwari, N. Sharma and H. Singh, "Predicting Ratings of Trending Youtube Videos using Machine Learning," 2022 5th International Conference on Contemporary Computing and Informatics (IC3I), Uttar Pradesh, India, 2022, pp. 27-31, doi: 10.1109/IC3I56241.2022.10072585.

[10] Sharma, R. (2024, April 15). *YouTube Trending Video Dataset (updated daily)*. Kaggle. https://www.kaggle.com/datasets/rsrishav/youtube-trending-video-dataset

[11] GeeksforGeeks, https://www.geeksforgeeks.org/ (accessed Nov. 2, 2024).